Multidimensional data clustering and dimension reduction for indexing and searching

Patent Number:	US6122628
Publication date:	2000-09-19
Inventor(s):	CASTELLI VITTORIO (US); THOMASIAN ALEXANDER (US); LI CHUNG-SHENG (US)
Applicant(s):	IBM (US)
Requested Patent:	JP11242674
Application Number:	US19970960540 19971031
Priority Number (s):	US19970960540 19971031
IPC Classification:	G06F17/30
EC Classification:	G06F17/30A
Equivalents:	CN1216841, DE69802960D, DE69802960T, EP1025514 (WO9923578), B1, HU0100581, JP3113861B2, PL340039, TW410304, WO9923578

Abstract

An improved multidimensional data indexing technique that generates compact indexes such that most or all of the index can reside in main memory at any time. During the clustering and dimensionality reduction, clustering information and dimensionality reduction information are generated for use in a subsequent search phase. The indexing technique can be effective even in the presence of variables which are not highly correlated. Other features provide for efficiently performing exact and nearest neighbor searches using the clustering information and dimensionality reduction information. One example of the dimensionality reduction uses a singular value decomposition technique. The method can also be recursively applied to each of the reduced-dimensionality clusters. The dimensionality reduction can also be applied to the entire database as a first step of the index generation.

Reference 1

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－242674

(43)公開日　平成11年(1999)９月７日

(51)Int.Cl.[6]	識別記号	ＦＩ		
G06F 17/30		G06F 15/401	310	D

審査請求　未請求　請求項の数55　ＯＬ　（全25頁）

(21)出願番号　特願平10－307241

(22)出願日　平成10年(1998)10月28日

(31)優先権主張番号　０８／９６０５４０

(32)優先日　1997年10月31日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレイション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国10504、ニューヨーク州アーモンク　（番地なし）

(72)発明者　ヴィットリオ・キャステリ
アメリカ合衆国10601、ニューヨーク州ホワイト・プレインズ、ノース・ブロードウェイ　55、アパートメント・２－17

(74)代理人　弁理士　坂口　博　（外２名）

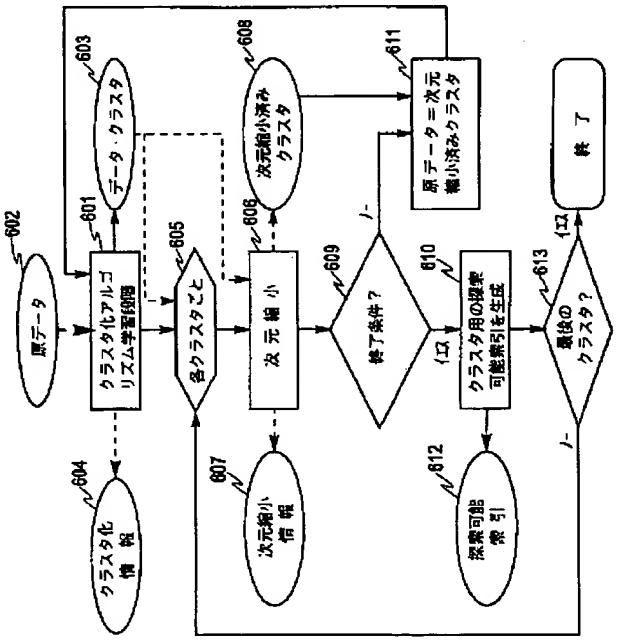
最終頁に続く

(54)【発明の名称】多次元データを表示する方法、プログラム記憶装置及び計算機用プログラム製品

(57)【要約】

【課題】多次元データのコンパクト表示を生成する装置及び方法を提供すること。

【解決手段】任意の時点で殆ど又は全ての索引が主メモリ内に常駐し得るように、コンパクト索引を生成するための改良された索引付け技法が開示されている。クラスタ化及び次元縮小ステップの間、後の探索段階で使用するのに備えて、クラスタ化情報及び次元縮小情報を生成する。この索引付け技法は、高度に相関していない変数の存在下においても、効率的に機能することができる。また、本発明は、クラスタ化情報及び次元縮小情報を使用して、絶対探索及び最近傍探索を効率的に行うことができる。次元縮小の１つの例は、特異値分解技法を使用する。また、この方法を、次元縮小済みクラスタの各々に再帰的に適用することができる。更に、次元縮小ステップを、索引生成の最初のステップとして、データベースの全体に適用することができる。



1

【特許請求の範囲】

【請求項１】多次元データを表示する方法であって、

（ａ）前記多次元データを１つ以上のクラスタに区分するステップと、（ｂ）前記１つ以上のクラスタに対するクラスタ化情報を生成及び記憶するステップと、（ｃ）前記１つ以上のクラスタに対する１つ以上の次元縮小済みクラスタ及び次元縮小情報を生成するステップと、

（ｄ）前記次元縮小情報を記憶するステップとを含んでいる、前記方法。

【請求項２】前記１つ以上の次元縮小済みクラスタに対する次元縮小済み索引を生成及び記憶するステップを更に含んでいる、請求項１記載の方法。

【請求項３】前記データが、複数のデータ・レコードを含んでいる空間データベース又はマルティメディア・データベース内に記憶され、

索引付けを遂行すべき前記データベースの表示を複数ベクトルの集合として作成するステップを更に含み、

前記各ベクトルが、前記データベース内の１行に対応し、前記各ベクトルの要素が、当該ベクトルに対応する特定の行の列内に保持されている値に対応し、これらの列について、探索可能な索引が生成され、

前記ステップ（ａ）が、前記ベクトルを１つ以上のクラスタに区分することを含んでいる、請求項１記載の方法。

【請求項４】前記索引を、全体として計算機の主メモリ内に記憶するステップを更に含んでいる、請求項２記載の方法。

【請求項５】前記ステップ（ｃ）が、特異値分解を含み、

前記各クラスタに対する変換行列及び当該変換行列の固有値を生成するステップと、

最大の固有値を含んでいる前記固有値の部分集合を選択するステップとを更に含み、

前記次元縮小情報が、前記変換行列及び前記固有値の部分集合を含んでいる、請求項２記載の方法。

【請求項６】前記次元縮小済み索引を使用して、指定データに最も類似するｋ個のレコードを探索するために、

前記記憶済みクラスタ化情報に基づいて、前記指定データを前記１つ以上のクラスタに関連付けるステップと、

前記関連付けられたクラスタに対する前記記憶済み次元縮小情報に基づいて、前記指定データを当該クラスタに対する部分空間に射影するステップと、

前記射影するステップに応答して、前記射影済み指定データに対する直交補空間を含んでいる次元縮小情報を生成するステップと、

前記索引を介して、前記射影済み指定データに最も類似するｋ個のレコードを有する前記関連付けられたクラスタを探索するステップと、

関連付けられた他の任意のクラスタが前記射影済み指定データに最も類似するｋ個のレコードのうち任意のレコードを含み得るか否かを決定するステップと、前記射影済み指定データに最も類似するｋ個のレコードのうち任意のレコードを含み得る前記任意のクラスタ上で、前記関連付けられたクラスタを探索するステップを反復するステップとを含んでいる、請求項５記載の方法。

2

【請求項７】前記索引を介して、前記関連付けられたクラスタを探索するステップが、

前記関連付けられたクラスタ内のｋ最近傍と前記射影済み指定データとの間の距離Ｄを、ミスマッチ索引δ^2の関数として、

δ^2（テンプレート，要素）＝D^2（射影済みテンプレート，要素）＋Σ‖直交補空間‖^2

の如く計算するステップを含んでいる、請求項６記載の方法。

【請求項８】前記指定データが、探索テンプレートを含み、

前記射影するステップが、前記次元縮小情報を使用して、前記テンプレートをそれが属するクラスタに関連する部分空間に射影するステップを含み、

前記射影済みテンプレートに対するテンプレート次元縮小情報を生成するステップを更に含み、

前記索引を介して、前記関連付けられたクラスタを探索するステップが、前記射影済みテンプレート及び前記テンプレート次元縮小情報に基づいて遂行され、

前記探索テンプレートに最も類似するｋ個のレコードのｋ最近傍集合を更新するステップを更に含んでいる、請求項６記載の方法。

【請求項９】前記固有値の部分集合を選択するステップが、戻される結果の精度及び再現度の関数である、請求項５記載の方法。

【請求項１０】指定データに最も類似するｋ個のレコードを探索するために、

前記クラスタ化情報に基づいて、前記指定データが属するクラスタを識別するステップと、

前記識別されたクラスタに対する前記次元縮小情報に基づいて、前記指定データの次元を縮小するステップと、

前記縮小するステップに応答して、次元縮小済み指定データに対する次元縮小情報を生成するステップと、

前記次元縮小情報を使用して、前記指定データが属するクラスタの次元縮小バージョンに対する多次元索引を探索するステップと、

前記多次元索引を介して、前記クラスタ内で前記最も類似するｋ個のレコードを検索するステップと、

前記検索済みの最も類似するｋ個のレコードのうち最も遠いレコードよりも前記指定データに近いレコードを保持し得る他のクラスタを識別するステップと、

前記識別するステップに応答して、前記指定データに最も近い他の候補クラスタを探索するステップと、

前記他の候補クラスタの全てについて前記他のクラスタを識別するステップ及び前記他の候補クラスタを探索す

るステップを反復するステップとを更に含んでいる、請求項２記載の方法。

【請求項１１】前記次元縮小バージョンのクラスタ内のｋ最近傍と前記射影済み指定データとの間の距離Ｄを、ミスマッチ索引δ^2の関数として、

δ^2（テンプレート，要素）＝D^2（射影済みテンプレート，要素）＋Σ‖直交補空間‖^2

の如く計算するステップを含んでいる、請求項１０記載の方法。

【請求項１２】前記クラスタ化情報が、前記１つ以上のクラスタの重心に関する情報を含み、

前記重心を一意的なラベルと関連付けるステップを更に含んでいる、請求項１記載の方法。

【請求項１３】前記データの次元が８より大きい、請求項１記載の方法。

【請求項１４】絶対探索を遂行するために、

前記記憶済みクラスタ化情報に基づいて、指定データを１つの前記クラスタに関連付けるステップと、

前記関連付けるステップに応答して、前記クラスタの次元縮小バージョンに対する記憶済み次元縮小情報に基づいて、前記指定データの次元を縮小するステップと、

前記次元縮小済み指定データに基づいて、前記指定データにマッチする前記クラスタの次元縮小バージョンを探索するステップとを更に含んでいる、請求項１記載の方法。

【請求項１５】前記探索するステップが、線形走査を遂行することを含んでいる、請求項１４記載の方法。

【請求項１６】前記ステップ（ａ）乃至（ｄ）を再帰的に適用することにより、前記次元縮小済みクラスタの階層を作成するステップと、

前記階層の最下位レベルにおけるクラスタに対する１つ以上の低次元索引を生成及び記憶するステップとを更に含んでいる、請求項１記載の方法。

【請求項１７】絶対探索を遂行するために、

（１）前記記憶済みクラスタ化情報を使用して、指定データが属するクラスタを探索すること、（２）前記記憶済み次元縮小情報を使用して、前記階層の対応する最下位レベルに到達するまで、前記指定データの次元を縮小すること、及び（３）前記低次元索引を使用して、前記指定データにマッチする前記クラスタの次元縮小バージョンを探索することを再帰的に適用するステップを更に含んでいる、請求項１６記載の方法。

【請求項１８】類似性に基づく探索を遂行するために、

（１）前記記憶済みクラスタ化情報を使用して、指定データが属する前記クラスタを探索すること、及び（２）前記記憶済み次元縮小情報を使用して、前記階層の最下位レベルに対応するように、前記指定データの次元を縮小することを再帰的に適用するステップと、

前記指定データが属する前記階層の最下位レベルにある終端クラスタから開始して、前記階層の各レベルにおいて、前記指定データのｋ最近傍のうち１つ以上のものを保持し得る候補終端クラスタを探索するステップと、

前記候補終端クラスタの各々ごとに、前記指定データに対するｋ最近傍についてクラスタ内探索を遂行するステップとを更に含んでいる、請求項１６記載の方法。

【請求項１９】類似性に基づく探索を遂行するために、

前記指定データの次元を縮小するステップと、

（１）前記記憶済みクラスタ化情報を使用して、次元縮小済み指定データが属する前記クラスタを探索すること、及び（２）前記記憶済み次元縮小情報を使用して、前記階層の最下位レベルに対応するように、前記次元縮小済み指定データの次元を縮小することを再帰的に適用するステップと、

前記指定データが属する前記階層の最下位レベルにある終端クラスタから開始して、前記階層の各レベルにおいて、前記次元縮小済み指定データのｋ最近傍のうち１つ以上のものを保持し得る候補終端クラスタを探索するステップと、

前記候補終端クラスタの各々ごとに、前記次元縮小済み指定データに対するｋ最近傍についてクラスタ内探索を遂行するステップとを更に含んでいる、請求項１６記載の方法。

【請求項２０】前記データが、データベース内に記憶され、

前記データベースの次元を縮小し且つ前記データベースに関連する次元縮小情報を生成するステップと、

前記データベースに関連する前記次元縮小情報を記憶するステップとを更に含み、

前記ステップ（ａ）が、前記次元縮小情報を生成するステップに応答して遂行される、請求項１記載の方法。

【請求項２１】絶対探索を遂行するために、

前記データベースに対する前記次元縮小情報に基づいて、指定データの次元を縮小するステップと、

前記縮小するステップに応答して、前記クラスタ化情報に基づいて、前記次元縮小済み指定データを１つの前記クラスタに関連付けるステップと、

前記関連付けられたクラスタに対する前記次元縮小情報に基づいて、前記指定データの次元を、関連付けられたクラスタによって定義される前記次元縮小済みクラスタの次元に縮小するステップと、

前記指定データの次元縮小済みバージョンに基づいて、これにマッチする次元縮小済みクラスタを探索するステップとを更に含んでいる、請求項２０記載の方法。

【請求項２２】類似性に基づく探索を遂行するために、

前記データベースに関連する前記次元縮小情報を使用して、指定データの次元を縮小するステップと、

前記クラスタ化情報を使用して、前記次元縮小済み指定データが属する前記クラスタを識別するステップと、

前記識別済みクラスタに対する前記次元縮小情報に基づいて、前記次元縮小済み指定データの次元を更に縮小す

るステップと、
次元を更に縮小した前記次元縮小済み指定データが属する前記クラスタの次元縮小バージョンを探索するステップと、
前記多次元索引を介して、前記クラスタ内の次元を更に縮小した前記次元縮小済み指定データに最も類似するｋ個のレコードを検索するステップと、
前記検索済みの最も類似するｋ個のレコードのうち最も遠いレコードよりも前記指定データに近いレコードを他のクラスタが保持し得るか否かを評価するステップと、
前記評価するステップに応答して、前記指定データに最も近い他のクラスタを探索するステップと、
前記他のクラスタの全てについて前記評価するステップ及び前記他のクラスタを探索するステップを反復するステップとを更に含んでいる、請求項２０記載の方法。
【請求項２３】前記データが、データベース内に記憶され、
前記１つ以上の次元縮小済みクラスタに対する１つ以上の次元縮小済み探索可能索引を生成及び記憶するステップを更に含んでいる、請求項２０記載の方法。
【請求項２４】絶対探索を遂行するために、
前記記憶済みクラスタ化情報に基づいて、指定データを１つの前記クラスタに関連付けるステップと、
前記関連付けるステップに応答して、前記指定データを、前記関連付けられたクラスタ及び前記関連付けられたクラスタに対する前記記憶済み次元縮小情報によって定義される次元縮小済みクラスタに分解するステップと、
前記分解済み指定データに基づいて、これにマッチする前記次元縮小済みクラスタに対する前記索引を探索するステップとを更に含んでいる、請求項２０記載の方法。
【請求項２５】前記指定データが、探索テンプレートを含み、
前記関連付けるステップが、前記記憶済みクラスタ化情報に基づいて、前記探索テンプレートが属するクラスタを識別することを含み、
前記分解するステップが、前記記憶済みクラスタ化情報に基づいて、前記探索テンプレートを前記識別済みクラスタに対する部分空間に射影することを含み、
前記探索するステップが、前記射影済みテンプレートについてクラスタ内探索を遂行することを含んでいる、請求項２４記載の方法。
【請求項２６】（ｅ）前記クラスタの形状に対する０次近似に対応するクラスタ境界を生成するステップと、
（ｆ）極小外接ボックスによって前記各クラスタの形状を近似し且つそれから前記各クラスタの形状に対する１次近似を生成するステップと、（ｇ）各次元の中点において前記外接ボックスを２ｋ個の超方形に区分するステップと、（ｈ）データ点を保持する超方形のみを保存し且つそれから前記クラスタの形状に対する２次近似を生成するステップと、（ｉ）前記保存済み超方形の各々ごとに、前記ステップ（ｇ）及び（ｈ）を反復することにより、前記クラスタに対する３次，４次，・・，ｎ次近似を逐次に生成するステップとを更に含んでいる、請求項１記載の方法。

【請求項２７】各クラスタの形状構造に対する逐次近似の階層を探索するために、
前記データベースに関連する前記次元縮小情報を使用して、前記指定データの次元を縮小するステップと、
前記クラスタ化情報に基づいて、前記次元縮小済み指定データが属するクラスタを識別するステップと、
前記識別済みクラスタに対する前記次元縮小情報に基づいて、前記次元縮小済み指定データの次元を更に縮小するステップと、
次元を更に縮小した前記次元縮小済み指定データが属するクラスタの次元縮小済みバージョンを探索するステップと、
前記多次元索引を介して、前記クラスタ内の次元を更に縮小した前記次元縮小済み指定データに最も類似するｋ個のレコードを検索するステップと、
前記検索済みの最も類似するｋ個のレコードのうち最も遠いレコードよりも前記指定データに近いレコードを１つ以上の他のクラスタが保持し得るか否かを評価するステップと、
前記クラスタの境界に基づいて、前記他のクラスタが前記指定データのｋ個のレコードのうち任意のレコードを保持し得る場合にのみ、当該他のクラスタを保存するステップと、
前記形状に対し次第に細密となる近似に基づいて、前記保存済みクラスタが前記ｋ最近傍のうち任意のものを保持し得るか否かを反復的に決定するとともに、当該クラスタが前記階層の最も細密なレベルにおいて受け入れられる場合にのみ前記保存済みクラスタを保存するステップと、
前記保存済みクラスタを保存するステップに応答して、前記保存済みクラスタを、前記データのｋ最近傍のうち１つ以上を保持する候補クラスタとして識別するステップとを更に含んでいる、請求項２６記載の方法。
【請求項２８】多次元データに対する１つ以上の次元縮小済み索引を含み、多次元データを表示する方法ステップを遂行するように計算機によって実行可能なプログラム命令を有形的に記憶しているプログラム記憶装置であって、
前記方法ステップが、（ａ）前記多次元データを１つ以上のクラスタに区分するステップと、（ｂ）前記１つ以上のクラスタに対するクラスタ化情報を生成及び記憶するステップと、（ｃ）前記１つ以上のクラスタに対する１つ以上の次元縮小済みクラスタ及び次元縮小情報を生成するステップと、（ｄ）前記次元縮小情報を記憶するステップとを含んでいる、前記プログラム記憶装置。

【請求項２９】前記１つ以上の次元縮小済みクラスタに対する次元縮小済み索引を生成及び記憶するステップを更に含んでいる、請求項２８記載のプログラム記憶装置。

【請求項３０】前記データが、複数のデータ・レコードを含んでいる空間データベース又はマルティメディア・データベース内に記憶され、

索引付けを遂行すべき前記データベースの表示を複数ベクトルの集合として作成するステップを更に含み、

前記各ベクトルが、前記データベース内の１行に対応し、前記各ベクトルの要素が、当該ベクトルに対応する特定の行の列内に保持されている値に対応し、これらの列について、探索可能な索引が生成され、

前記ステップ（ａ）が、前記ベクトルを１つ以上のクラスタに区分することを含んでいる、請求項２８記載のプログラム記憶装置。

【請求項３１】前記索引を、全体として計算機の主メモリ内に記憶するステップを更に含んでいる、請求項２９記載のプログラム記憶装置。

【請求項３２】前記ステップ（ｃ）が、特異値分解を含み、

前記各クラスタに対する変換行列及び当該変換行列の固有値を生成するステップと、

最大の固有値を含んでいる前記固有値の部分集合を選択するステップとを更に含み、

前記次元縮小情報が、前記変換行列及び前記固有値の部分集合を含んでいる、請求項２９記載のプログラム記憶装置。

【請求項３３】前記次元縮小済み索引を使用して、指定データに最も類似するｋ個のレコードを探索するために、

前記記憶済みクラスタ化情報に基づいて、前記指定データを前記１つ以上のクラスタに関連付けるステップと、

前記関連付けられたクラスタに対する前記記憶済み次元縮小情報に基づいて、前記指定データを当該クラスタに対する部分空間に射影するステップと、

前記射影するステップに応答して、前記射影済み指定データに対する直交補空間を含んでいる次元縮小情報を生成するステップと、

前記索引を介して、前記射影済み指定データに最も類似するｋ個のレコードを有する前記関連付けられたクラスタを探索するステップと、

関連付けられた他の任意のクラスタが前記射影済み指定データに最も類似するｋ個のレコードのうち任意のレコードを含み得るか否かを決定するステップと、

前記射影済み指定データに最も類似するｋ個のレコードのうち任意のレコードを含み得る前記任意のクラスタ上で、前記関連付けられたクラスタを探索するステップを反復するステップとを含んでいる、請求項３２記載のプログラム記憶装置。

【請求項３４】前記索引を介して、前記関連付けられたクラスタを探索するステップが、

前記関連付けられたクラスタ内のｋ最近傍と前記射影済み指定データとの間の距離Ｄを、ミスマッチ索引δ^2の関数として、

δ^2（テンプレート，要素）＝D^2（射影済みテンプレート，要素）＋Σ‖直交補空間‖^2

の如く計算するステップを含んでいる、請求項３３記載のプログラム記憶装置。

【請求項３５】前記指定データが、探索テンプレートを含み、

前記射影するステップが、前記次元縮小情報を使用して、前記テンプレートをそれが属するクラスタに関連する部分空間に射影するステップを含み、

前記射影済みテンプレートに対するテンプレート次元縮小情報を生成するステップを更に含み、

前記索引を介して、前記関連付けられたクラスタを探索するステップが、前記射影済みテンプレート及び前記テンプレート次元縮小情報に基づいて遂行され、

前記探索テンプレートに最も類似するｋ個のレコードのｋ最近傍集合を更新するステップを更に含んでいる、請求項３３記載のプログラム記憶装置。

【請求項３６】前記固有値の部分集合を選択するステップが、戻される結果の精度及び再現度の関数である、請求項３２記載のプログラム記憶装置。

【請求項３７】指定データに最も類似するｋ個のレコードを探索するために、

前記クラスタ化情報に基づいて、前記指定データが属するクラスタを識別するステップと、

前記識別されたクラスタに対する前記次元縮小情報に基づいて、前記指定データの次元を縮小するステップと、

前記縮小するステップに応答して、次元縮小済み指定データに対する次元縮小情報を生成するステップと、

前記次元縮小情報を使用して、前記指定データが属するクラスタの次元縮小バージョンに対する多次元索引を探索するステップと、

前記多次元索引を介して、前記クラスタ内で前記最も類似するｋ個のレコードを検索するステップと、

前記検索済みの最も類似するｋ個のレコードのうち最も遠いレコードよりも前記指定データに近いレコードを保持し得る他のクラスタを識別するステップと、

前記識別するステップに応答して、前記指定データに最も近い他の候補クラスタを探索するステップと、

前記他の候補クラスタの全てについて前記他のクラスタを識別するステップ及び前記他の候補クラスタを探索するステップを反復するステップとを更に含んでいる、請求項２９記載のプログラム記憶装置。

【請求項３８】前記次元縮小バージョンのクラスタ内のｋ最近傍と前記射影済み指定データとの間の距離Ｄを、ミスマッチ索引δ^2の関数として、

9

δ^2（テンプレート，要素）＝D^2（射影済みテンプレート，要素）＋Σ‖直交補空間‖^2

の如く計算するステップを含んでいる、請求項３７記載のプログラム記憶装置。

【請求項３９】前記クラスタ化情報が、前記１つ以上のクラスタの重心に関する情報を含み、

前記重心を一意的なラベルと関連付けるステップを更に含んでいる、請求項２８記載のプログラム記憶装置。

【請求項４０】前記データの次元が８より大きい、請求項２８記載のプログラム記憶装置。

【請求項４１】絶対探索を遂行するために、

前記記憶済みクラスタ化情報に基づいて、指定データを１つの前記クラスタに関連付けるステップと、

前記関連付けるステップに応答して、前記クラスタの次元縮小バージョンに対する記憶済み次元縮小情報に基づいて、前記指定データの次元を縮小するステップと、

前記次元縮小済み指定データに基づいて、前記指定データにマッチする前記クラスタの次元縮小バージョンを探索するステップとを更に含んでいる、請求項２８記載のプログラム記憶装置。

【請求項４２】前記探索するステップが、線形走査を遂行することを含んでいる、請求項４１記載のプログラム記憶装置。

【請求項４３】前記ステップ（ａ）乃至（ｄ）を再帰的に適用することにより、前記次元縮小済みクラスタの階層を作成するステップと、

前記階層の最下位レベルにおけるクラスタに対する１つ以上の低次元索引を生成及び記憶するステップとを更に含んでいる、請求項２８記載のプログラム記憶装置。

【請求項４４】絶対探索を遂行するために、

（１）前記記憶済みクラスタ化情報を使用して、指定データが属するクラスタを探索すること、（２）前記記憶済み次元縮小情報を使用して、前記階層の対応する最下位レベルに到達するまで、前記指定データの次元を縮小すること、及び（３）前記低次元索引を使用して、前記指定データにマッチする前記クラスタの次元縮小バージョンを探索することを再帰的に適用するステップを更に含んでいる、請求項４３記載のプログラム記憶装置。

【請求項４５】類似性に基づく探索を遂行するために、

（１）前記記憶済みクラスタ化情報を使用して、指定データが属する前記クラスタを探索すること、及び（２）前記記憶済み次元縮小情報を使用して、前記階層の最下位レベルに対応するように、前記指定データの次元を縮小することを再帰的に適用するステップと、

前記指定データが属する前記階層の最下位レベルにある終端クラスタから開始して、前記階層の各レベルにおいて、前記指定データのｋ最近傍のうち１つ以上のものを保持し得る候補終端クラスタを探索するステップと、

前記候補終端クラスタの各々ごとに、前記指定データに対するｋ最近傍についてクラスタ内探索を遂行するステップとを更に含んでいる、請求項４３記載のプログラム記憶装置。

10

【請求項４６】類似性に基づく探索を遂行するために、

前記指定データの次元を縮小するステップと、

（１）前記記憶済みクラスタ化情報を使用して、次元縮小済み指定データが属する前記クラスタを探索すること、及び（２）前記記憶済み次元縮小情報を使用して、前記階層の最下位レベルに対応するように、前記次元縮小済み指定データの次元を縮小することを再帰的に適用するステップと、

前記指定データが属する前記階層の最下位レベルにある終端クラスタから開始して、前記階層の各レベルにおいて、前記次元縮小済み指定データのｋ最近傍のうち１つ以上のものを保持し得る候補終端クラスタを探索するステップと、

前記候補終端クラスタの各々ごとに、前記次元縮小済み指定データに対するｋ最近傍についてクラスタ内探索を遂行するステップとを更に含んでいる、請求項４３記載のプログラム記憶装置。

【請求項４７】前記データが、データベース内に記憶され、

前記データベースの次元を縮小し且つ前記データベースに関連する次元縮小情報を生成するステップと、

前記データベースに関連する前記次元縮小情報を記憶するステップとを更に含み、

前記ステップ（ａ）が、前記次元縮小情報を生成するステップに応答して遂行される、請求項２８記載のプログラム記憶装置。

【請求項４８】絶対探索を遂行するために、

前記データベースに対する前記次元縮小情報に基づいて、指定データの次元を縮小するステップと、

前記縮小するステップに応答して、前記クラスタ化情報に基づいて、次元縮小済み指定データを１つの前記クラスタに関連付けるステップと、

前記関連付けられたクラスタに対する前記次元縮小情報に基づいて、前記指定データの次元を、関連付けられたクラスタによって定義される前記次元縮小済みクラスタの次元に縮小するステップと、

前記指定データの次元縮小済みバージョンに基づいて、これにマッチする次元縮小済みクラスタを探索するステップとを更に含んでいる、請求項４７記載のプログラム記憶装置。

【請求項４９】類似性に基づく探索を遂行するために、

前記データベースに関連する前記次元縮小情報を使用して、指定データの次元を縮小するステップと、

前記クラスタ化情報を使用して、前記次元縮小済み指定データが属する前記クラスタを識別するステップと、

前記識別済みクラスタに対する前記次元縮小情報に基づいて、前記次元縮小済み指定データの次元を更に縮小するステップと、

次元を更に縮小した前記次元縮小済み指定データが属する前記クラスタの次元縮小バージョンを探索するステップと、
前記多次元索引を介して、前記クラスタ内の次元を更に縮小した前記次元縮小済み指定データに最も類似するｋ個のレコードを検索するステップと、
前記検索済みの最も類似するｋ個のレコードのうち最も遠いレコードよりも前記指定データに近いレコードを他のクラスタが保持し得るか否かを評価するステップと、
前記評価するステップに応答して、前記指定データに最も近い他のクラスタを探索するステップと、
前記他のクラスタの全てについて前記評価するステップ及び前記他のクラスタを探索するステップを反復するステップとを更に含んでいる、請求項４７記載のプログラム記憶装置。
【請求項５０】前記データが、データベース内に記憶され、
前記１つ以上の前記次元縮小済みクラスタに対する１つ以上の次元縮小済み探索可能索引を生成及び記憶するステップを更に含んでいる、請求項４７記載のプログラム記憶装置。
【請求項５１】絶対探索を遂行するために、
前記記憶済みクラスタ化情報に基づいて、指定データを１つの前記クラスタに関連付けるステップと、
前記関連付けるステップに応答して、前記指定データを、前記関連付けられたクラスタ及び当該関連付けられたクラスタに対する記憶済み次元縮小情報によって定義される前記次元縮小済みクラスタに分解するステップと、
前記分解済み指定データに基づいて、これにマッチする前記次元縮小済みクラスタに対する前記索引を探索するステップとを更に含んでいる、請求項４７記載のプログラム記憶装置。
【請求項５２】前記指定データが、探索テンプレートを含み、
前記関連付けるステップが、前記記憶済みクラスタ化情報に基づいて、前記探索テンプレートが属するクラスタを識別することを含み、
前記分解するステップが、前記記憶済みクラスタ化情報に基づいて、前記探索テンプレートを前記識別済みクラスタに対する部分空間に射影することを含み、
前記探索するステップが、前記射影済みテンプレートについてクラスタ内探索を遂行することを含んでいる、請求項５１記載のプログラム記憶装置。
【請求項５３】（ｅ）前記クラスタの形状に対する０次近似に対応するクラスタ境界を生成するステップと、
（ｆ）極小外接ボックスによって各クラスタの形状を近似し且つそれから各クラスタの形状に対する１次近似を生成するステップと、（ｇ）各次元の中点において前記外接ボックスを２ｋ個の超方形に区分するステップと、
（ｈ）データ点を保持する超方形のみを保存し且つそれから前記クラスタの形状に対する２次近似を生成するステップと、（ｉ）前記保存済み超方形の各々ごとに、前記ステップ（ｇ）及び（ｈ）を反復することにより、前記クラスタに対する３次，４次，・・，ｎ次近似を逐次に生成するステップとを更に含んでいる、請求項２８記載のプログラム記憶装置。
【請求項５４】各クラスタの形状構造に対する逐次近似の階層を探索するために、
前記データベースに関連する前記次元縮小情報を使用して、前記指定データの次元を縮小するステップと、
前記クラスタ化情報に基づいて、前記次元縮小済み指定データが属するクラスタを識別するステップと、
前記識別済みクラスタに対する前記次元縮小情報に基づいて、前記次元縮小済み指定データの次元を更に縮小するステップと、
次元を更に縮小した前記次元縮小済み指定データが属するクラスタの次元縮小済みバージョンを探索するステップと、
前記多次元索引を介して、前記クラスタ内の次元を更に縮小した前記次元縮小済み指定データに最も類似するｋ個のレコードを検索するステップと、
前記検索済みの最も類似するｋ個のレコードのうち最も遠いレコードよりも前記指定データに近いレコードを１つ以上の他のクラスタが保持し得るか否かを評価するステップと、
前記クラスタの境界に基づいて、前記他のクラスタが前記指定データのｋ個のレコードのうち任意のレコードを保持し得る場合にのみ、当該他のクラスタを保存するステップと、
前記形状に対し次第に細密となる近似に基づいて、前記保存済みクラスタが前記ｋ最近傍のうち任意のものを保持し得るか否かを反復的に決定するとともに、当該クラスタが前記階層の最も細密なレベルにおいて受け入れられる場合にのみ前記保存済みクラスタを保存するステップと、
前記保存済みクラスタを保存するステップに応答して、前記保存済みクラスタを、前記データのｋ最近傍のうち１つ以上を保持する候補クラスタとして識別するステップとを更に含んでいる、請求項５３記載のプログラム記憶装置。
【請求項５５】多次元データを表示するための計算機可読プログラム・コード手段を有形的に記憶する記憶媒体を備えた計算機用プログラム製品であって、
前記計算機可読プログラム・コード手段が、
前記多次元データを１つ以上のクラスタに区分するように計算機を制御するためのクラスタ化手段と、
前記クラスタ化手段に結合され、前記１つ以上のクラスタに対するクラスタ化情報を生成及び記憶するように計算機を制御するための手段と、

前記クラスタ化手段に結合され、前記１つ以上のクラスタに対する１つ以上の次元縮小済みクラスタ及び次元縮小情報を生成するように計算機を制御するための次元縮小手段と、

前記次元縮小手段に結合され、前記次元縮小情報を記憶するように計算機を制御するための手段と、

前記１つ以上の次元縮小済みクラスタに対する次元縮小済み索引を生成及び記憶するように計算機を制御するための手段を含んでいる、前記計算機用プログラム製品。

【発明の詳細な説明】

【０００１】

【関連する出願】本発明の関連出願は、"Searching Multidimensional Indexes Using Associated Clustering and Dimension Reduction Information"と題する、１９９７年１０月３１日に出願した係属中の米国特許出願（本出願人の整理番号ＹＯ９９７３６８）である。本明細書では、この関連出願の内容を援用する。

【０００２】

【発明の属する技術分野】本発明は、多次元データのコンパクト表示を生成及び探索することに係り、更に詳細に説明すれば、関連付けられたクラスタ化情報及び次元縮小情報を使用して、データベース・システム内にある多次元データのコンパクト索引表示を生成及び探索することに係る。

【０００３】

【従来の技術】多次元索引付けは、空間データベースにとって基本的な技法であり、地図情報システム（ＧＩＳ）や、膨大なデータ・ウェアハウスを使用して意思決定を支援するためのオンライン分析処理（ＯＬＡＰ）や、イメージ及びビデオ・データから高次元の特徴ベクトルを抽出するようなマルチメディア・データベースに対し広範に適用することができる。

【０００４】意思決定支援システムは、事業体が成功するための不可欠の技術になりつつある。各事業体は、意思決定支援システムを使用して、商用データベースから（データ・ウェアハウスとも呼ばれる）有用な情報を推論することができる。商用データベースが状態情報を保持するのに対し、データ・ウェアハウスは履歴情報を保持するのが普通である。一般に、データ・ウェアハウスのユーザは、個々のレコードを別々に調べるのではなく、傾向を把握することに関心を持っている。従って、意思決定用の照会を主体とする計算を集中的に遂行して、集約機能を多量に使用する。その結果、計算が完了するまでの遅れが長くなって、生産性上の制約が受け入れがたいものになることがある。

【０００５】かかる遅れを減少させるのに使用されていた公知の技法は、出現頻度が高い照会を予め計算するか、又はサンプリング技法を使用する、というものである。特に、最近の注目を集めているのは、データ・キューブのようなオンライン分析処理（ＯＬＡＰ）技法を、

大規模な関係データベース又は意思決定支援用のデータ・ウェアハウスに適用することである（例えば、Jim Gray et al, "Data Cube:A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, andSub-Totals", International Conference on Data Engineering, 1996, New Orleans, pp. 152-160 を参照）。ここでは、ユーザは、データ・ウェアハウスからの履歴データを多次元データ・キューブとして観察するのが普通である。このキューブ内の各セル（又は格子点）は、総売上高のような諸項目の集約から成るビューである。

【０００６】多次元索引付けは、異なるタイプの照会に回答するために使用することができる。これらの照会を例示すると、次の通りである。

（１）索引付けした列の指定値を有するレコードを探索する（絶対探索）。

（２）［ａ１，・・ａ２］、［ｂ１，・・ｂ２］，・・、［ｚ１，・・ｚ２］内のレコードを探索する（範囲探索）。但し、ａ、ｂ及びｚは、異なる次元を表すものとする。

（３）ユーザが指定したテンプレート又は例に対し最も類似するｋ個のレコードを探索する（ｋ最近傍探索）。

【０００７】多次元索引付けは、イメージ・マイニングにも適用することができる。イメージ・マイニング製品の一例は、「ＭＥＤＩＡＭＩＮＥＲ」（本出願人の商標）である。この製品が提供する２つのツール、すなわち「イメージ・コンテンツ照会プログラム（ＱＢＩＣ）」及び「ＩＭＡＧＥＭＩＮＥＲ」は、手動的に作成した関連キーワードのリストを使用して探索を遂行するのではなく、コンテンツを分析することによってイメージの検索を遂行する。

【０００８】ＱＢＩＣを使用するのに適したアプリケーションは、キーワードを使用しても適切な結果が得られないようなもの、例えば博物館及び美術館用のライブラリに関係するアプリケーション、又は電子商取引用のオンライン在庫品の写真に関係するアプリケーションである。ちなみに、後者のアプリケーションでは、顧客がビジュアル・カタログを使用し、必要とする商品（例えば、壁紙や衣服）のカラーやテクスチャなどを吟味した上で、所望の商品を探索することができる。

【０００９】「ＩＭＡＧＥＭＩＮＥＲ」のようなイメージ・マイニング用アプリケーションは、概念的な照会（例えば、「森の景色」、「氷」、「シリンダ」など）を使用して、イメージ・データベースに対する照会を遂行することができる。カラーやテクスチャ及び輪郭などのイメージ・コンテンツは、システムが自動的に認識可能な単純オブジェクトとして結合される。

【００１０】これらの単純オブジェクトは、知識ベース内に表示される。これを分析して得られるテキストの記述は、後の検索に備えて索引付けされる。

【００１１】データベースの照会を実行中、データベー

ス探索プログラムは、記憶データの一部及び索引付け構造の一部をアクセスする。この場合、アクセスされるデータ量は、照会のタイプ、ユーザが提供するデータ及び索引付けアルゴリズムの効率に依存する。大規模データベースでは、そのデータ及び索引付け構造の少なくとも一部を、計算機システムのメモリ階層のうち１台以上のハードディスクから成る大容量の低速部分に記憶するのが普通である。探索プロセスの間、かかるデータ及び索引付け構造の一部を、メモリ階層の高速部分、すなわち主メモリ及びキャッシュ・メモリの１つ以上のレベルにロードする。一般に、メモリ階層の高速部分は、比較的高価であるという理由で、メモリ階層の記憶容量のうち数パーセントに相当するにすぎない。キャッシュ・メモリの１つ以上のレベルに完全にロード可能な命令及びデータを使用するプログラムは、主メモリ内に存在する命令及びデータを使用するプロセスよりも高速及び効率的であり、また後者のプロセスは、ハードディスク上に存在する命令及びデータを使用するプログラムよりも高速である。この場合の技術的制限とは、主メモリ及びキャッシュ・メモリのコストが比較的高いために、大規模データベースを完全に記憶するのに十分な主メモリ及びキャッシュ・メモリを備えた計算機システムを構築することが実際的でない、というものである。

【００１２】従って、当分野において要請されている索引付け技法とは、任意の時点で殆ど又は全ての索引を主メモリ内に常駐可能にするようなサイズの索引を生成するとともに、探索プロセス中にハードディスクから主メモリに転送すべきデータの量を制限するようなものである。本発明は、かかる要請を実現することに向けられている。

【００１３】Ｒツリーを含む周知の空間索引付け技法は、これを範囲照会及び最近傍照会のために使用することができる。Ｒツリーの詳細は、例えば A. Gutman, "R-trees: A Dynamic Index Structure for Spatial Searching", ACM SIGMOD Conf. onManagement of Data, Boston, MA, June 1994 に記述されている。しかしながら、これらの技法は、特徴空間の次元の数が増大するにつれて、探索空間が次第に疎になるために、その効率が急速に低下する。例えば、次元の数が８よりも大きい場合、Ｒツリーのような方法は有用でないことが知られている。この場合の有用基準は、要求を完了するための時間と、データベース内の全てのレコードを順次に走査してこの要求を完了するという暴力的戦略が必要とする時間との比較によって決まる。高次元空間における通常の索引付け技法の非効率性は、「次元の呪い」と呼ばれる現象に起因する。この現象については、例えば V. Cherkassky et al, "From Statistics to Neural Networks", NATO ASI Series, Vol. 136, Springer-Verlag, 1994 に記述されている。また、この現象に起因して、高次元の特徴空間については、索引空間を複数の超立方体

にクラスタ化することが非効率的となる。

【００１４】前述のように、高次元の特徴空間を索引付けするための既存の空間索引付け技法が非効率的であることから、周知の技法は、特徴空間の次元の数を縮小することに向けられている。例えば、次元の数を縮小するために、（特徴選択とも呼ばれる）可変部分集合選択を使用するか、又は特異値分解の後に可変部分集合選択を使用することができる（例えば、C. T. Chen, "Linear System Theory and Design", Holt, Rinechart and Winston, Appendix E, 1984 を参照）。可変部分集合選択は、統計分野で盛んに研究されていて、多数の方法が既に提案されている（例えば、 Shibata et al, "An Optimal Selection of Regression Variables", Biometrika, Vol. 68, No. 1, 1981, pp. 45-54 を参照）。これらの方法が索引生成システムにおいて有効であるのは、多くの変数（データベース内の列）が高度に相関している場合だけである。

【００１５】従って、高度に相関していない変数の存在下であっても、高次元データの索引付けを効率的に遂行するような改良された索引付け技法が要請されている。この技法は、メモリの利用度及び探索速度の観点から、索引を効率的に生成しなければならない。本発明は、これらの要請を実現することに向けられている。

【００１６】

【発明が解決しようとする課題】前述の要請に従って、本発明の目的は、多次元データのコンパクト表示を生成するための装置及び方法を提供することにある。本発明は、データベース用の探索可能な多次元索引を生成することを特徴としている。また、本発明は、これらの索引を柔軟に生成するとともに、絶対探索及び類似性に基づく探索を効率的に遂行することを特徴としている。更に、本発明は、探索プロセス中にハードディスクから主メモリに転送するデータの量を制限することを可能にする、コンパクト索引を生成することを特徴としている。

【００１７】本発明の１つの適用例は、多次元索引付けである。多次元索引付けは、空間データベースにとって基本的な技法であり、地図情報システム（ＧＩＳ）や、膨大なデータ・ウェアハウスを使用して意思決定を支援するためのオンライン分析処理（ＯＬＡＰ）や、イメージ及びビデオ・データから高次元の特徴ベクトルを抽出するマルチメディア・データベースのマイニングを遂行するためのイメージ・マイニング製品に対し広範に適用することができる。

【００１８】

【課題を解決するための手段】本発明の特徴を有する方法の一例は、（ａ）多次元データを１つ以上のクラスタに区分するステップと、（ｂ）前記クラスタに対するクラスタ化情報を生成及び記憶するステップと、（ｃ）前記クラスタに対する１つ以上の次元縮小済みクラスタ及び次元縮小情報を生成するステップと、（ｄ）前記次元

縮小情報を記憶するステップとを含んでいる。また、本発明は、前記次元縮小済みクラスタに対する次元縮小済み索引を生成及び記憶することを特徴としている。

【００１９】各クラスタ内で使用されている索引付け技法に依存して、対応する索引付け機構を使用することにより、目標ベクトルを検索することができる。例えば、各クラスタ内の索引付けのために、通常の多次元空間索引付け技法（Ｒツリーを含むがこれに限らない）を使用することができる。代替的に、どんな空間索引付け構造も利用できない場合は、クラスタ内探索機構として、暴力的又は線形走査を利用することができる。

【００２０】推奨実施例において、ステップ（ｃ）は、特異値分解であり、かくて分解済みの指定データに基づいて、次元縮小済みクラスタにマッチする索引を探索する。次元縮小情報の一例は、（固有値及び固有ベクトルを含む）変換行列である。この変換行列は、特異値分解及びこの変換行列の選択された固有値によって生成される。

【００２１】本発明の特徴に従った多次元索引を生成する方法の他の例は、索引付けを遂行すべきデータベースの表示を、複数ベクトルの集合として作成するステップを含んでいる。但し、各ベクトルは、このデータベース内の１行に対応し、各ベクトルの要素は、このベクトルに対応する特定の行の列（そのための索引を生成しなければならない）内に保持されている値に対応する。この方法では、クラスタ化技法を使用して、このベクトル集合を（クラスタとも呼ばれる）１つ以上のグループに区分するとともに、このクラスタ化に関連するクラスタ化情報を生成及び記憶する。次いで、次元縮小情報を使用して、これらのグループの各々に対し次元縮小技法を別個に適用することにより、クラスタ内にある諸要素の低次元表示を生成する。次に、次元縮小済みクラスタの各々ごとに、このクラスタの次元の数について効率的な索引を生成する技法を使用して、索引を生成する。

【００２２】本発明の他の特徴に従って、この方法は、これが再帰的となるように、次元縮小済みクラスタの各々に対し別個に適用することができる。次元縮小ステップ及びクラスタ化ステップの両者を再帰的に適用するというこのプロセスが終了するのは、もはや次元を縮小させることができないような場合である。

【００２３】本発明の他の特徴に従って、次元縮小ステップは、（データベースを区分するステップの前に）索引生成プロセスにおける最初のステップとして、データベースの全体に適用することができる。（クラスタ化とも呼ばれる）区分ステップ及び次元縮小ステップの間、探索段階で使用するのに備えて、クラスタ化情報及び次元縮小情報を生成する。

【００２４】本発明の更に他の特徴に従って、次元縮小ステップを容易にするように、クラスタ化ステップを適当に選択することができる。例えば、これを遂行するには、ユークリッド距離のような空間的に不変の距離関数から導かれる損失を最小にするのではなく、データの局所共分散構造に従って空間を区分するというクラスタ化方法を使用することができる。

【００２５】また、本発明は、指定データに最も類似するｋ個の要素のうち最も遠い要素よりも指定データに近い要素を、他のクラスタが保持し得るか否かを評価することを特徴としている。公知のように、クラスタ化情報を使用すると、複数の区分の境界を再構成することが可能であり、またこれらの境界を使用すると、指定データに最も類似するｋ個の要素のうち１つの要素を、１つのクラスタが保持し得るか否か決定することができる。当業者には明らかなように、これらのクラスタ境界は、クラスタ自体の構造に対する簡単な近似である。すなわち、この境界の数学的形式からは、この境界上の所与の位置にクラスタの要素が存在するか否かを断定することはできない。一例として、データベースが２つの球型データ・クラスタを保持しており、そしてこれらのクラスタが互いに著しく離れているようなケースを検討する。このケースの妥当な境界は、これらのクラスタの重心を結合する線分に垂直で且つこれらの重心から等距離の超平面となろう。これらのクラスタは互いに著しく離れているから、この境界近傍には如何なるデータ点も存在しない。他のケースでは、この境界は、両クラスタの多数の要素に非常に接近することがあり得る。従って、本発明は、各クラスタの実際の形状構造に対する諸近似の階層を使用することにより、このクラスタが、指定データに最も類似するｋ個の要素のうち１つ以上の要素を保持し得るか否かを決定することを特徴としている。

【００２６】推奨実施例において、本発明は、計算機可読プログラム記憶媒体上に有形的に記憶したソフトウェアとして実現することができる。このソフトウェアを構成するプログラム命令を計算機によって実行すると、多次元データを表示する次の方法、すなわち多次元データのコンパクト表示を生成するステップと、データベース用の探索可能な多次元索引を生成するステップと、これらの索引を使用して絶対探索及び類似性に基づく探索を効率的に遂行するステップとを含む方法が実施される。

【００２７】

【発明の実施の形態】図１には、本発明の特徴を実現したネットワーク化クライアント／サーバ・システムが例示されている。図示のように、複数クライアント１０１及び複数サーバ１０６は、ネットワーク１０２によって相互接続されている。サーバ１０６は、データベース管理システム（ＤＢＭＳ）１０４及び直接アクセス記憶装置（ＤＡＳＤ）１０５を含んでいる。一般に、クライアント１０１上で照会を生成し、これをネットワーク１０２を通してサーバ１０６に発信する。かかる照会は、ユーザが提供する例又は探索テンプレートのような指定デ

19

ータを含んでおり、ＤＡＳＤ１０５内に記憶されているデータベースを検索又は更新を遂行するためにＤＢＭＳ１０４と対話する。ＤＢＭＳ１０４の一例は、ＩＢＭ社から提供されている「ＤＢ２」（本出願人の商標）である。

【００２８】本発明の１つの側面に従って、多次元の、例えば空間索引付けを必要とする照会（範囲照会及び最近傍照会を含む）は、多次元索引付けエンジン１０７を呼び出す。この多次元索引付けエンジン１０７（図８～図１１を参照して後述）は、本発明の索引生成論理１１０（図６及び図７を参照して後述）が生成する１つ以上のコンパクト多次元索引１０８、クラスタ化情報１１１及び次元縮小情報１１２に基づいて、この照会が指定する制約を満足するようなベクトル又はレコードを検索する。本発明のコンパクト索引１０８の全部又は殆どは、サーバ１０６の主メモリ及びキャッシュ・メモリの少なくとも一方に記憶することが好ましい。空間データベースのようなデータベースは、１つ以上のシステム上に存在することができる。また、多次元索引付けエンジン１０７及び索引生成論理１１０の少なくとも一方は、ＤＢＭＳ１０４の一部として統合化することができる。更に、（探索論理とも呼ばれる）多次元索引付けエンジン１０７及び索引生成論理１１０は、サーバ１０６上で実行可能な計算機プログラム製品上のソフトウェアとして有形的に実施することができる。

【００２９】１つの適用例は、スーパーマーケットの記憶済み販売時点情報管理トランザクションであり、店の位置（緯度及び経度）の幾何座標を含むものである。ここでは、サーバ１０６は、記憶済みデータから知識又はパターンを探索するために意思決定支援タイプのアプリケーションをサポートしていることが好ましい。例えば、オンライン分析処理（ＯＬＡＰ）エンジン１０３を使用して、ＯＬＡＰに関係する照会をインターセプトすることにより、これらの処理を容易に遂行することができる。本発明に従って、ＯＬＡＰエンジン１０３は、可能であればＤＢＭＳ１０４と関連して、多次元索引エンジン１０７を使用することにより、ＯＬＡＰに関係する照会用の索引１０８を探索する。なお、本発明の索引生成論理１１０は、データ・ウェアハウスの多次元データ・キューブ表示にも適用することができる。データ・ウェアハウスの多次元データ・キューブ表示を生成するための方法及びシステムは、"System and Method for Generating Multi-Representation of a Data Cube"と題する、１９９７年４月１４日に出願された係属中の米国特許出願第８４３２９０号に記述されている。本明細書では、この出願の内容を援用する。

【００３０】空間索引付けの効果を享受するデータの他の例は、マルチメディア・データである。オーディオ、ビデオ及びイメージのようなマルチメディア・データは、索引付けのために使用するメタデータとは別個に記

20

憶することができる。メディア・データの索引付け及び検索を容易にするために使用可能なメタデータの１つの重要な成分は、生データから抽出されるような「特徴ベクトル」である。例えば、イメージの領域からテクスチャ、カラー・ヒストグラム及び形状を抽出し、これを検索用の索引１０８を構成するために使用することができる。

【００３１】イメージ・マイニング・アプリケーションの例は、ＩＢＭ社の「DB2 Image Extender」における統合化探索機構のＱＢＩＣである。ＱＢＩＣは、イメージ照会エンジン（サーバ）、並びにＨＴＭＬグラフィカル・ユーザ・インタフェース及び関連する共通ゲートウェイ・インタフェース（ＣＧＩ）スクリプトから成るサンプル・クライアントを含んでおり、これらが相まって完全なアプリケーションの基礎を形成している。このサーバ及びクライアントの両者は、ユーザがアプリケーションに特有のイメージ・マッチング関数を開発し且つこれをＱＢＩＣに追加することができるように、拡張可能となっている。また、イメージ探索サーバは、ビジュアル・イメージのコンテンツに基づいて、大規模イメージ・データベースの照会を可能にする。

【００３２】この特徴は、次の通りである。

（１）ビジュアル・メディアの形式で照会すること。例えば、「これと同様のイメージを示せ」という具合。但し、「同様」の意味を、カラーや、レイアウト、テクスチャなどで定義する必要がある。

（２）照会イメージに対する類似性に従って、イメージをランク付けすること。

（３）諸イメージの自動的索引付けを遂行すること。但し、カラー及びテクスチャの数値記述子を記憶する必要がある。探索の間、類似イメージを探索するために、これらのプロパティを使用する。

（４）ビジュアル照会とテキスト照会又は日付のようなパラメータに関する照会との組み合わせ。

【００３３】同様に、索引付けを遂行すべきデータベースの表示を複数ベクトルの集合として作成することにより、複数の索引を生成することができる。但し、各ベクトルは、このデータベース内の１行に対応し、各ベクトルの要素は、この特定の行について、それぞれの列（そのための索引を生成することが必要である）に保持されている値に対応する。

【００３４】データベースの表示を複数ベクトルの集合として作成することは、当分野では周知である。この表示を作成するための代表的な方法は、データベースの各行ごとに、生成すべき索引の次元に等しい長さを有する１つの配列を作成し、この配列の諸要素に対し、この配列に対応するデータベースの特定の行の複数の列（そのための索引を生成しなければならない）に保持されている値をコピーする、というものである。

【００３５】ここで、ベクトルｖのｉ番目の要素がv_i

であると仮定すると、ベクトルｖを次のように表すことができる。

【００３６】

【数１】

$$v = [v_1 \ldots v_N] \qquad (1)$$

【００３７】但し、Ｎは、索引付けのために使用するベクトルの次元の数である。

【００３８】一般に、クライアント側は、３つのタイプの照会を指定することができる。これらの照会の全ては、所定の形式の空間索引付けを必要とする（従来技術の項の説明を参照）。これらの照会は、次の通りである。

（１）絶対照会：１つのベクトルを指定すると、このベクトルにマッチするレコード又はマルチメディア・データが検索される。

（２）範囲照会：このベクトルの各次元の下限及び上限が指定される。

（３）最近傍照会：類似性の測度に基づいて、最も「類似する」ベクトルが検索される。

【００３９】２つのベクトルｖ１及びｖ２の間の最も普通に使用されている類似性の測度は、ユークリッド距離の測度ｄであり、これは次のように定義されている。

【００４０】

【数２】

$$d^2 = \Sigma (v1_i - v2_i)^2 \qquad (2)$$

【００４１】ここで、次元ｉの全てが範囲照会又は最近傍照会の計算に必ずしも関与する必要はないことに留意されたい。両ケースにおいて、結果を検索するために、これらの次元の部分集合を指定することができる。

【００４２】図２には、多次元空間における複数のベクトルの分布が例示されている。図示のように、この空間の全体を表示するには、３次元が必要である。しかしながら、ｘ－ｙ平面、ｙ－ｚ平面及びｚ－ｘ平面上に位置するクラスタ２０１、２０２及び２０３の各々を表示するには、２次元だけが必要である。従って、かかるデータの適正なクラスタ化を通して、次元縮小を達成可能であると結論することができる。これと同じ次元縮小は、特異値分解だけでは達成することができない。なぜなら、特異値分解は、特徴空間内の軸が主要な次元（この例では３）とマッチするように、特徴空間を再配向するにすぎないからである。

【００４３】１つのベクトルの１つ以上の次元を除去することは、元の点を部分空間に射影することと等価である。式（２）は、このベクトル内の個々の要素が異なっているような次元だけを計算すればよいことを示している。その結果、このベクトルを部分空間に射影しても、距離の計算には影響しないから、除去される要素は元の空間内で変化しないことになる。

【００４４】図３には、元の３次元空間内の３点を２次元部分空間に射影するに際し、これらの３点のうち任意の２点間の相対距離を維持するようにした距離計算の一例が示されている。図示のように、元の３次元空間では、点３０１と点３０２との間の距離は、点３０１と点３０３との間の距離よりも大きい。ここで、これらの点を２次元部分空間に射影した結果である点（３０４、３０５、３０６）間の相対距離が維持されていることに留意されたい。

【００４５】図４には、３次元空間内の３点を２次元部分空間に射影するに際し、相対距離のランクに影響を及ぼすようにした一例が示されている。図示のように、３次元空間内の点４０１と点４０２との間の距離は、点４０２と点４０３との間の距離よりも大きい。しかしながら、このケースでは、射影済みの点４０４と点４０５との間の距離は、射影済みの点４０５と点４０６との間の距離よりも小さい。従って、射影済みの２次元部分空間では、２点間の相対距離が維持されないことになる。

【００４６】以下では、複数のベクトルを部分空間に射影する際に生じ得るような最大誤差を評価するための技法を導出する。先ず、このプロセスは、最大誤差の上界を決定することから開始する。１つのクラスタの重心Ｖｃを、次のように定義する。

【００４７】

【数３】

$$V_c = \frac{1}{N} \Sigma V_i \qquad (3)$$

【００４８】但し、Ｎは、このクラスタ内にある複数のベクトル｛Ｖ１，・・，ＶＮ｝の総数である。このクラスタをｋ次元の部分空間に射影した後、そこで一般性を失わないように最後の（ｎ－ｋ）次元を除去するものとすると、元の空間に比較して、この部分空間における任意の２ベクトル間の距離に対する誤差が生ずることになる。この誤差の項は、次のように表される。

【００４９】

【数４】

$$Error^2 = \Sigma_{i=k+1}^{n} (V_{1,i} - V_{2,i})^2 \qquad (4)$$

【００５０】式（４）から、次の不等式が直ちに成立する。

【００５１】

【数５】

$$Error^2 \le \sum_{i=k+1}^{n} (|V_{1,i}| + |V_{2,i}|)^2$$
$$\le \sum_{i=k+1}^{n} (2\max(|V_{1,i}|, |V_{2,i}|))^2$$
$$\le 4 \sum_{i=k+1}^{n} \max(|V_{1,i}|, |V_{2,i}|)^2 \qquad (5)$$

【００５２】式（５）は、射影済みの部分空間における距離を計算する際に生ずる最大誤差が束縛されていることを示している。

【００５３】図５には、本発明に従った距離の計算を遂行する際の近似の一例が示されている。テンプレート点Ｔ（５０１）と生成点Ｖ（５０６）との間の距離は、前掲の式（２）によって与えられる。このユークリッド距離は、基準座標系の回転、座標系の原点の平行移動、座標軸の鏡映及び座標の順序付けに関し、不変である。そこで、一般性を失わないように、Ｖ（５０６）がクラスタ１（５０５）に属するものとする。次に、クラスタ１（５０５）の共分散行列の固有ベクトルによって定義される基準座標系を考慮し、この基準座標系の原点が重心１であるものとする。そうすると、Ｔ（５０１）とクラスタ１（５０５）内にあるＶ（５０６）との間の距離は、次のように表すことができる。

【００５４】

【数６】

$$d^2 = \sum_{i=1}^{n} (T_i - V_i)^2 \qquad (6)$$

【００５５】但し、座標T_i及びV_iは、基準座標系に対し相対的である。次に、最後のｎ－ｋ＋１個の座標をゼロに設定することにより、クラスタ１（５０５）を部分空間１に射影する。かくて、Ｔ（５０１）とＶ（５０６）を部分空間１に射影した点Ｖ’（５０７）との間の距離は、次のように表すことができる。

【００５６】

【数７】

$$d'^2 = \sum_{i=1}^{k} (T_i - V_i)^2 + \sum_{i=k+1}^{n} (T_i)^2$$
$$= d_1^2 + d_2^2 \qquad (7)$$

【００５７】項d_1は、部分空間１へのＴ（５０１）の射影５４０、すなわち射影１と、部分空間１へのＶ（５０６）の射影Ｖ’（５０７）との間のユークリッド距離である。項d_2は、Ｔ（５０１）と、部分空間１へのその射影、すなわち射影１との間の距離である。換言すれば、項d_2は、Ｔ（５０１）と部分空間１との間の距離である。かくて、Ｔ（５０１）とベクトルＶ（５０６）との間の距離を計算する際に、式（６）に式（７）を代入すると、導かれる近似を束縛することができる。初等幾何学が教えるところによれば、検討中の３点、すなわちＴ（５０１）、Ｖ（５０６）及びＶ’（５０７）は、一意的な２次元部分空間（平面）を識別する。説明を簡潔にするため、この平面が図５の平面５２０に対応するものとする。そうすると、式（６）で定義されている距離ｄは、Ｔ（５０１）及びＶ（５０６）を結合する線分の長さに等しくなり、式（７）で定義されている距離ｄ’は、Ｔ（５０１）及びＶ’（５０７）を結合する線分の長さに等しくなる。初等幾何学の周知の定理によれば、３角形の１辺の長さは、他の２辺の長さの差の絶対値よりも長く、それらの和よりも短い。このことが暗示するのは、式（６）で定義されている距離ｄを式（７）で定義されている距離ｄ’で置換する際に生ずる誤差が、Ｖ（５０６）及びＶ’（５０７）を結合する線分の長さよりも短いか又はこれに等しい、ということである。従って、自乗誤差は、次のように束縛されることになる。

【００５８】

【数８】

$$error^2 \le \Sigma_{i=k+1}^{n} (V_i)^2 \qquad (8)$$

【００５９】図６には、次元縮小済みクラスタの階層及びこの階層の最下部にあるクラスタ用の低次元索引を生成するための流れ図が示されている。ステップ６０１で、クラスタ化プロセスは、原データ６０２を入力として受け取り、このデータを複数のデータ・クラスタ６０３に区分するとともに、この区分の詳細に関するクラスタ化情報６０４を生成する。原データ６０２内の各エントリは、式（１）で定義されているようなベクトル属性を含んでいる。クラスタ化アルゴリズムは、当分野で公知のクラスタ化又はベクトル量子化アルゴリズムのうち任意のものを選択することができる。これらのアルゴリズムについては、例えば Leonard Kauffman et al, "Finding Groups in Data", John Wiley & Sons, 1990、又は YosephLinde et al, "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications, Vol. COM-28, No. 1, January 1980, pp. 84-95 を参照されたい。クラスタ化アルゴリズムの学習段階によって生ぜられるクラスタ化情報６０４は、このアルゴ

25

リズムとともに変化する。すなわち、クラスタ化情報６０４は、クラスタ化アルゴリズムの分類段階によって、新しい任意の、以前に見られなかったサンプルを生成するとともに、各クラスタごとに１つの表現ベクトルを生成することを可能にする。クラスタ化情報６０４は、各々が一意的なラベルに関連する、複数のクラスタの重心を含んでいることが好ましい（例えば、前掲の Yoseph Linde et al, "An Algorithm for Vector Quantizer Design" を参照）。ステップ６０５で、順序づけ論理が開始して、後続の動作が各クラスタに対し個別的に且つ順次的に適用されるように、動作の流れを制御する。なお、複数の計算回路が設けられている場合には、順序付け論理６０５をディスパッチング論理によって置き換えることにより、各々が異なるデータ・クラスタについて動作する、複数の計算回路上で同時的な計算を行わせることができる。ステップ６０６で、次元縮小論理６０６は、１つのデータ・クラスタ６０３を受け取り、次元縮小情報６０７及び次元縮小済みデータ・クラスタ６０８を生成する。ステップ６０９で、終了条件のテストを遂行する（後述）。もし、終了条件が満足されなければ、ステップ６１１で、原データ６０２を次元縮小済みデータ・クラスタ６０８で置換した後、このプロセスをステップ６０１に戻すことにより、ステップ６０１～６０９を再帰的に適用することができる。他方、終了条件が満足されると、ステップ６１０で、現クラスタ用の探索可能索引６１２を生成する。もし、ステップ６１３で、既に分析済みのクラスタの数が、ステップ６０１でクラスタ化アルゴリズムによって生成したクラスタ６０３の数と等しいことが分かれば、このプロセスが終了する。さもなければ、このプロセスはステップ６０５に戻る。一般に、クラスタ６０３の数は、ユーザによって選択されるが、当分野では自動式手順も知られている（例えば、Brian Everitt, "Cluster Analysis", Halsted Press, 1973, Chapter 4.2 を参照）。

【００６０】ステップ６０９における終了条件のテストは、次のように定義されるデータ・ボリュームＶ（Ｘ）の概念を基礎とすることができる。

【００６１】

【数９】

$$V(X) = \Sigma_{i=1}^{m} n_i \qquad (9)$$

【００６２】但し、Ｘは複数レコードの集合であり、n_i はｉ番目のレコードであり、和はＸの全ての要素についてのものである。もし、これらのレコードが次元縮小ステップ６０６の前に同じサイズＳを有し、そしてｎが１つのクラスタ内にあるレコードの数を表すのであれば、このクラスタのボリュームはＳｎとなる。他方、Ｓ’が次元縮小ステップ６０６の後のレコードのサイズを表すものとすれば、次元縮小後のこのクラスタのボリュームはＳ’ｎとなる。終了条件は、ボリュームＳｎ及びＳ’ｎを比較してテストすることができ、Ｓｎ＝Ｓ’ｎとなる場合にこのプロセスが終了する。

26

【００６３】他の実施例では、ステップ６０９における終了条件のテストが存在しないので、このプロセスの再帰的適用は遂行されない。

【００６４】図７には、ステップ６０６の次元縮小論理が例示されている。図示のように、ステップ７０１で、特異値分解論理は、データ・クラスタ７０２を入力として受け取り、１つの変換行列７０３及びその複数の固有値７０４を生成する。変換行列７０３の複数の列は、この行列の複数の固有ベクトルである。特異値分解用のアルゴリズムは、当分野では周知である（例えば、R. A. Horn et al, "MatrixAnalysis", Cambridge University Press (1985) を参照）。当業者には明らかなように、代替実施例では、ステップ７０１の特異値分解論理を、当分野では周知の主成分分析論理によって置き換えることができる。

【００６５】ステップ７０５で、分類論理は、固有値７０４を入力として受け取り、大きさの減少順に分類済みの固有値７０６を生成する。かかる分類論理は、当分野では周知の任意の分類論理とすることができる。ステップ７０７で、選択論理は、所定の選択基準に従って、最大の固有値７０８を保持する処の順序付けられた固有値の部分集合７０６を選択する。かかる選択基準の一例は、その和が変換行列７０３のトレースのユーザ指定パーセントよりも大きい、最小グループの固有値を選択するようなものである。当分野では周知のように、このトレースは、対角線上の要素の和である。この例では、変換行列７０３及び選択済み固有値７０８が、次元縮小情報６０７を構成する。代替的に、固有値の選択を、精度と再現度のトレードオフに基づいて遂行することができる（後述）。

【００６６】ステップ７０９で、変換論理は、データ・クラスタ７０２及び変換行列７０３を入力として受け取り、変換行列７０３が指定する変換をデータ・クラスタ７０２の諸要素に適用することによって、変換済みデータ・クラスタ７１０を生成する。ステップ７１１で、選択済み固有値７０８及び変換済みデータ・クラスタ７１０を使用して、次元縮小済みデータ・クラスタ７１２を生成する。推奨実施例では、最小数の次元を保存することにより、次元縮小を遂行している。このように最小数の次元を保存するのは、対応する固有値の集合が全体的な分散のうち少なくとも一定のパーセント（例えば、９５％）を占めるようにするためである。

【００６７】代替的に、固有値の選択を、精度と再現度のトレードオフに基づいて遂行することもできる。精度と再現度を理解するには、本発明の１つの方法によって遂行される探索動作が、（図１０及び図１１を参照して後述するように）近似を対象とし得ることに留意する必要がある。ここで、ｋを、Ｎ個の要素を有するデータベ

27

ースにおいて１つのテンプレートに対し最も類似する要素の所望の数であるものとする（ｋ最近傍）。この探索動作は近似を対象とするものであるから、ユーザは、ｋよりも多い数の結果が戻されることを要求するのが普通である。ｎを、戻される結果の数であるものとすると、ｎ個の結果のうちｃ個の結果だけが正しい。つまり、ｃ個の結果が、このテンプレートに対するｋ最近傍に含まれている、ということである。精度は、戻された結果と正しい結果の比であって、これを次のように定義する。

精度　＝ｃ／ｎ

再現度は、探索動作によって戻された正しい結果の比であって、これを次のように定義する。

再現度＝ｃ／ｋ

精度及び再現度はテンプレートの選択に応じて変化するから、それらの予期値は、当該システムの性能の比較的良好な測度となる。かくて、精度及び再現度の両者は、複数のテンプレートの分布にわたって、固定値ｎ及びｋの関数として取られた予期値（Ｅ）を意味する。

精度　＝Ｅ（ｃ）／ｎ

再現度＝Ｅ（ｃ）／ｋ

明らかに、戻される結果の数ｎが増大するにつれて、精度が減少するのに対し、再現度は増大する。精度及び再現度の傾向は、単調ではないのが普通である。Ｅ（ｃ）はｎに依存するから、効率－再現度の曲線は、ｎのパラメータ関数としてプロットされることが多い。推奨実施例では、要求元が、探索の所望の精度及び許容再現度に関する下限を指定する。その場合、次元縮小論理（図６のステップ６０６）は、精度及び再現度に基づいて、次元縮小を遂行する。すなわち、複数の固有値を減少順に順序付けた後、次元縮小論理は、最小の固有値に対応する次元を除去し、そして原トレーニング集合又はユーザが提供するトレーニング集合から無作為に選択したテスト・サンプル集合に基づいて、結果的な精度－再現度関数を推定する。次元縮小論理は、この精度－再現度関数から、所望の再現度が得られるような精度の最大値ｎ$_{max}$を導出する。引き続いて、次元縮小論理は、次の最小の固有値を除去して同じ手順を反復し、所望の再現度が得られるような対応する精度を計算する。かかる反復手順が終了するのは、計算済みの精度がユーザ指定のスレッショルド値よりも小さくなる場合である。その場合、次元縮小論理は、かかる終了条件が生ずる直前の反復手順で保存されていた次元だけを保存するのである。

【００６８】本発明の他の実施例では、要求元が、所望の再現度の値だけを指定するのに対し、次元縮小論理は、所望の再現度を得るために精度を増大させる場合のコストを推定する。このコストは、２つの成分を有する。一方の成分は、低次元の空間における距離の計算及び最近傍探索が一層効率的であるという理由で、次元の数に応じて減少する。他方の成分は、所望の再現度を保証するために、保存済み次元の数が減少するにつれて検索済み結果の数が増大しなければならないという理由で、次元の数に応じて増大する。効率的な方法を使用したとしても、比較的大きな数ｎの最近傍を検索することは高価につく。というのは、分析しなければならない探索空間の部分が、所望の結果の数に応じて増大するからである。その場合、次元縮小論理は、全数探索を遂行することにより、ユーザが指定した再現度の値のための探索コストを最小限にする処の、保存すべき次元の数を探索する。

【００６９】クラスタ化及び特異値分解ステップは、終了条件に到達するまで（ステップ６０９）、複数のベクトルに対し再帰的に適用することができる（ステップ６０１～６１１）。かかる終了条件の１つは、本明細書に記述されているように、各クラスタの次元がもはや縮小不能となるような場合である。オプションとして、Ｒツリーのような通常の空間索引付け技法を各クラスタに適用することもできる。これらの技法は、次元を最小化したクラスタについては一層効率的である。こうして、高次元ベクトルの集合についての索引生成プロセスが完了する。

【００７０】図８～図１５を参照して後述するように、本発明は、多次元データのコンパクト表示を使用して効率的な探索を遂行することをも特徴としている。当業者には明らかなように、本発明の探索方法は、本明細書に記述されている多次元データの特定のコンパクト表示に限られるものではない。

【００７１】図８には、本発明に従って生成した探索可能な多次元索引（図１の１０８又は図６の６１２）に基づいて、絶対探索を遂行するための論理の流れが例示されている。この例では、索引を生成するに当たり、クラスタ化及び特異値分解ステップを再帰的に適用していない。絶対探索とは、探索照会（例えば、探索テンプレート）に正確にマッチする、１つ以上のレコードを検索するためのプロセスである。図示のように、ステップ８０２で、（クラスタ探索論理とも呼ばれる）多次元索引エンジン１０７が、探索テンプレート８０１のような指定データを含んでいる１つの照会を入力として受け取る。ステップ８０２で、図６のステップ６０１で生成したクラスタ化情報６０４を使用して、探索テンプレート８０１が属するクラスタを識別する。ステップ８０３で、図６のステップ６０６で生成した次元縮小情報６０７を使用して、探索テンプレート８０１をステップ８０２で識別したクラスタに関連する部分空間に射影することにより、射影済みテンプレート８０４を生成する。ステップ８０３で、クラスタ内探索論理は、図６のステップ６１０で生成した探索可能な多次元索引６１２を使用して、射影済みテンプレート８０４についての探索を遂行する。空間索引付け構造が利用不能である場合、最も簡単なクラスタ内の探索機構は、線形走査（又は線形探索）を遂行するというものである。クラスタの次元が比較的

小さな場合（例えば、１０より小さな場合）、Ｒツリーのような空間索引付け構造は、線形走査よりも良好な効率を有するのが普通である。

【００７２】図９には、本発明に従って生成した探索可能な多次元索引（図１の１０８又は図６の６１２）に基づいて、絶対探索を遂行するための他の論理の流れが例示されている。この例における索引（１０８又は６１２）は、クラスタ化及び次元縮小論理を再帰的に適用して生成されたものである。絶対探索とは、探索照会に正確にマッチする１つ以上のレコードを検索するためのプロセスである。図示のように、ステップ９０２で、（図８のステップ８０２におけるクラスタ探索論理に類似する）クラスタ探索論理が、探索テンプレート９０１のような指定データを含んでいる１つの照会を入力として受け取る。ステップ９０２で、図６のステップ６０１で生成したクラスタ化情報６０４を使用して、探索テンプレート９０１が属するクラスタを識別する。（図８のステップ８０３に類似する）ステップ９０３で、図６のステップ６０６で生成した次元縮小情報６０７を使用して、探索テンプレート９０１をステップ９０２で識別したクラスタに関連する部分空間に射影することにより、射影済みテンプレート９０４を生成する。ステップ８０５で、現クラスタが終端であるか否か、すなわち多次元索引構成プロセスの間に、現クラスタに対しそれ以上の再帰的なクラスタ化及び特異値分解ステップが適用されなかったか否かを決定する。もし、現クラスタが終端でなければ、ステップ９０７で、探索テンプレート９０１を射影済みテンプレート９０４によって置き換えた後、このプロセスはステップ９０２に戻る。他方、現クラスタが終端であれば、ステップ９０６で、クラスタ内探索論理は、探索可能な多次元索引６１２を使用して、射影済みテンプレート９０４についての探索を遂行する。前述のように、空間索引付け構造が利用不能である場合、最も簡単なクラスタ内探索機構は、線形走査（又は線形探索）を遂行するというものである。クラスタの次元が比較的小さな場合（例えば、１０より小さな場合）、Ｒツリーのような空間索引付け構造は、線形走査よりも良好な効率を有するのが普通である。

【００７３】また、本発明は、指定データに最も類似する検索済みのｋ個の要素のうち最も遠い要素よりも指定データに近いような要素を他のクラスタが保持し得るか否かを評価することを特徴としている。公知のように、クラスタ化情報を使用して複数の区分の境界を再構成することが可能であり、またこれらの境界を使用して１つのクラスタがｋ最近傍を保持し得るか否か決定することができる。当業者には明らかなように、これらのクラスタ境界は、クラスタ自体の構造に対する簡単な近似である。換言すれば、この境界の数学的形式からは、この境界上の所与の位置にクラスタの要素が存在するか否かを断定することはできない。一例として、データベースが

２つの球形データ・クラスタを保持しており、そしてこれらのクラスタが互いに著しく離れているようなケースを検討する。このケースの妥当な境界は、これらのクラスタの重心を結合する線分に垂直で且つこれらの重心から等距離の超平面となろう。しかしながら、これらのクラスタは互いに著しく離れているから、この境界近傍には如何なるデータ点も存在しないことになる。他のケースでは、この境界は、両クラスタの多数の要素に非常に接近することがあり得る。

【００７４】図１４及び図１５を参照して後述するように、本発明は、クラスタ化情報に加えて、各クラスタの実際の形状構造に対する近似階層を計算及び記憶するとともに、かかる近似階層を使用することにより、所与のベクトルからの一定の距離よりも近いような要素を保持し得るクラスタを識別することを特徴としている。

【００７５】図１０には、本発明に従って生成した探索可能な多次元索引（図６の６１２）に基づいて、ｋ最近傍探索プロセスを遂行するための論理の流れが例示されている。この例では、索引を生成するに当たり、クラスタ化及び特異値分解ステップを再帰的に適用していない。ｋ最近傍探索とは、探索テンプレートの形式を有する指定データに最も類似する、データベース内のｋ個のエントリを戻すためのプロセスである。ステップ１００２で、所望のマッチの数に等しいｋ（１０００）を使用して、ｋ最近傍集合１００９を初期化する。その目的は、ｋ最近傍集合１００９が高々ｋ個の要素を保持し、しかも次のステップが開始する前に空であるようにすることである。ステップ１００３で、クラスタ探索論理は、探索テンプレート１００１のような照会を入力として受け取るとともに、図６のステップ６０１で生成したクラスタ化情報６０４を使用して、探索テンプレート１００１が属するクラスタを識別する。ステップ１００４で、次元縮小情報６０７を使用して、探索テンプレート１００１をステップ１００３で識別したクラスタに関連する部分空間に射影する。この射影ステップ１００４は、射影済みテンプレート１００６及び次元縮小情報１００５を生成する。後者の次元縮小情報１００５は、射影済みテンプレート１００６の直交補空間（探索テンプレート１００１及び射影済みテンプレート１００６のベクトル差によって定義されるもの）及びこの直交補空間のユークリッド距離を含んでいる。ステップ１００７で、クラスタ内探索論理は、次元縮小情報１００５及び射影済みテンプレート１００６に加え、探索可能な多次元索引６１２を使用して、ｋ最近傍集合１００９を更新することができる。本発明に従って適応可能なクラスタ内探索論理は、公知の最近傍探索方法のうち任意のものとすることができる（例えば、公知の方法については、"Nearest Neighbor Pattern Classification", B. V. Desarathy (editor), IEEE Computer Society (1991) を参照）。本発明に従ったクラスタ内探索論理（ス

31

テップ１００７）は、例えば、射影済みテンプレート１００６と縮小済み次元を有するベクトル空間内にあるクラスタの諸メンバとの間の自乗距離を計算するステップと、その結果を探索テンプレート１００１とこのクラスタの部分空間との間の自乗距離に加えるステップと、その最終的な結果を直交補空間の自乗長さの「和」として定義するステップとを含んでいる。この論理の結果は、次元縮小情報１００５の一部として、ステップ１００４で次のように計算される。

δ^2（テンプレート，要素）＝D^2（射影済みテンプレート，要素）＋Σ‖直交補空間‖^2。

【００７６】もし、ステップ１００７の開始時に、ｋ最近傍集合１００９が空であれば、クラスタ内探索論理は、現クラスタ内にある要素の数がｋより大きいときは、射影済みテンプレート１００６に最も近い現クラスタのｋ個の要素で、さもなければ現クラスタの全ての要素で、ｋ最近傍集合１００９を充填してこれを更新する。ｋ最近傍集合１００９の各要素は、対応するミスマッチ索引δ^2に関連付けられている。

【００７７】他方、ステップ１００７の開始時に、ｋ最近傍集合１００９が空でなければ、クラスタ内探索論理は、ｋ最近傍集合１００９内の要素に現に関連する最大の索引よりも小さなミスマッチ索引δ^2を有するような要素が探索されるときに、ｋ最近傍集合１００９を更新する。ｋ最近傍集合１００９を更新するためには、そこから最大のミスマッチ索引δ^2を有する要素を除去するとともに、これを新しく探索された要素で置き換えることができる。

【００７８】もし、ｋ最近傍集合１００９がｋ個よりも少ない要素を保持していれば、不在の要素は、無限の距離にある要素と見なされる。ステップ１００８で、最近傍を保持し得る他の候補クラスタの存否を決定する。このステップは、クラスタ化情報６０４を入力として受け取り、これからクラスタ境界を決定することができる。もし、探索テンプレート１００１が属さないようなクラスタの境界が、ｋ最近傍集合１００９の最も遠い要素よりも近ければ、このクラスタは候補クラスタである。もし、候補クラスタが全く存在しなければ、このプロセスは終了し、ｋ最近傍集合１００９の内容が結果として戻される。さもなければ、このプロセスはステップ１００４に戻り、そこで現クラスタがステップ１００８で識別した候補クラスタとなる。

【００７９】図１１には、本発明に従って生成した探索可能な多次元索引（図６の６１２）に基づいて、ｋ最近傍探索プロセスを遂行するための論理の流れが例示されている。この例では、索引を生成するに当たり、クラスタ化及び特異値分解ステップを再帰的に適用している。ｋ最近傍探索とは、探索テンプレートの形式を有する指定データに最も類似する、データベース内のｋ個のエントリを戻すためのプロセスである。ステップ１１０２

32

で、ｋ最近傍集合１１１１を空に初期化し、そしてこれが高々ｋ個の要素を保持し得るように、所望のマッチの数ｋ（１１００）を使用して、ｋ最近傍集合１１１１を初期化する。ステップ１１０３で、クラスタ探索論理は、探索テンプレート１１０１を入力として受け取るとともに、図６のステップ６０１で生成したクラスタ化情報６０４を使用して、探索テンプレート１１０１を対応するクラスタに関連付ける。ステップ１１０４で、図６のステップ６０６で生成した次元縮小情報６０７を使用し、探索テンプレート１１０１をステップ１１０３で識別したクラスタに関連付けられた部分空間に射影する。この射影ステップ１１０４は、射影済みテンプレート１１０６及び次元縮小情報１１０５を生成する。後者の次元縮小情報１１０５は、射影済みテンプレート１１０６の直交補空間（探索テンプレート１１０１及び射影済みテンプレート１１０６のベクトル差によって定義されるもの）及びこの直交補空間のユークリッド距離を含んでいることが好ましい。ステップ１１０７で、現クラスタが終端であるか否か、すなわち索引を構成する間に、現クラスタに対しそれ以上の再帰的なクラスタ化及び特異値分解ステップが適用されなかったか否かを決定する。もし、現クラスタが終端でなければ、ステップ１１０８で、探索テンプレート１１０１を射影済みテンプレート１１０６によって置き換えた後、このプロセスはステップ１１０３に戻る。さもなければ、ステップ１１０９で、クラスタ内探索論理は、次元縮小情報１１０５及び射影済みテンプレート１１０６に加え、探索可能な多次元索引６１２を使用して、ｋ最近傍集合１１１１を更新する。本発明に従って適応可能なクラスタ内探索論理は、当分野で公知の最近傍探索方法のうち任意のものとすることができる（例えば、前掲の"Nearest Neighbor Pattern Classification", B. V. Desarathy (editor), IEEE Computer Society (1991) を参照）。本発明に従ったクラスタ内探索論理（ステップ１１０９）は、例えば、射影済みテンプレート１１０６と縮小済み次元を有するベクトル空間内にある現クラスタの諸メンバとの間の自乗距離を計算するステップと、その結果を探索テンプレート１１０１と現クラスタの部分空間との間の自乗距離に加えるステップと、その最終的な結果を直交補空間の自乗長さの「和」として定義するステップとを含んでいる。この論理の結果は、次元縮小情報１１０５の一部として、ステップ１１０４で次のように計算される。

δ^2（テンプレート，要素）＝D^2（射影済みテンプレート，要素）＋Σ‖直交補空間‖^2。

【００８０】もし、ステップ１１０９の開始時に、ｋ最近傍集合１１１１が空であれば、クラスタ内探索論理は、現クラスタ内にある要素の数がｋ個よりも大きいときは、射影済みテンプレート１１０６に最も近い現クラスタのｋ個の要素で、或いは現クラスタ内にある要素の

数がｋに等しいか又はこれより小さいときは、現クラスタの全ての要素で、ｋ最近傍集合１１１１を充填してこれを更新する。ｋ最近傍集合１１１１の各要素は、その対応するミスマッチ索引δ^2に関連付けられていることが好ましい。

【００８１】もし、ステップ１１０９の開始時に、ｋ最近傍集合１１１１が空でなければ、クラスタ内探索論理は、ｋ最近傍集合１１１１内の要素に現に関連する最大の索引よりも小さなミスマッチ索引δ^2を有するような要素が探索されるときに、ｋ最近傍集合１１１１を更新する。ｋ最近傍集合１１１１を更新するためには、そこから最大のミスマッチ索引δ^2を有する要素を除去するとともに、これを新しく探索された要素で置き換えればよい。

【００８２】もし、ｋ最近傍集合１１１１がｋ個よりも少ない要素を保持していれば、不在の要素は、無限の距離にある要素と見なされる。ステップ１１１０で、（最初のクラスタ化ステップを適用する前に）階層の現レベルが、その最上レベルであるか否かを決定する。もし、現レベルが最上レベルであれば、このプロセスは終了し、ｋ最近傍集合１１１１の内容が結果として戻される。他方、現レベルが最上レベルでなければ、ステップ１１１４で、現レベルにおける候補クラスタ、すなわちｋ最近傍のうち幾つかのものを保持し得るクラスタを探索する。この探索は、次元縮小情報１１０５及びクラスタ化情報６０４を使用して遂行される。ステップ１１１４で、クラスタ化情報６０４を使用して、クラスタ境界を決定する。もし、探索テンプレート１１０１が属さないようなクラスタの境界が、ｋ最近傍集合１１１１の最も遠い要素よりも近ければ、このクラスタは候補クラスタである。もし、候補クラスタが全く存在しなければ、ステップ１１１３で、現レベルを階層の先行レベルに設定し、次元縮小情報１１０５を更新した後、このプロセスはステップ１１１０に戻る。他方、候補クラスタが存在すれば、ステップ１１１５で、テンプレートを候補クラスタに射影して射影済みテンプレート１１０６を更新するとともに、次元縮小情報１１０５を更新する。次いで、このプロセスはステップ１１０７に戻る。

【００８３】図１２の（ａ）～（ｃ）は、類似性のみに基づくクラスタ化技法の結果を、データの局所構造に適応させたアルゴリズムを使用したクラスタ化の結果と比較している。類似性のみに基づくクラスタ化技法は、例えば、各クラスタの諸要素と対応する重心との間のユークリッド距離を最小化することを基礎としている（前掲の Linde et al, "An Algorithm for Vector Quantizer Design" を参照）。詳述すると、図１２の（ａ）は、基準座標系１２０１及び複数ベクトルの集合１２０２を示している。もし、各クラスタの諸要素と対応する重心との間のユークリッド距離を最小化することを基礎とするクラスタ化技法が使用されているのであれば、その可

能な結果は、図１２の（ｂ）に示されている通りである。すなわち、ベクトル集合１２０２は、超平面１２０３及び１２０４によって、クラスタ１（１２０５）、クラスタ２（１２０６）及びクラスタ３（１２０７）から成る、３つのクラスタに区分される。かかる結果的なクラスタは、互いに類似するベクトルを保持しているが、データの局所構造を捕捉しないから、準－最適の次元縮小を与える。図１２の（ｃ）は、データの局所構造に適応させたアルゴリズムを使用したクラスタ化の結果を示している。このクラスタ化から得られる３つのクラスタ、すなわちクラスタ１（１２０８）、クラスタ２（１２０９）及びクラスタ３（１２１０）は、データの局所構造を良好に捕捉するから、独立の次元縮小に対し適応させることができる。

【００８４】図１３には、データの局所構造に適応させたクラスタ化アルゴリズムが例示されている。ステップ１３０２で、クラスタ化すべきベクトル集合１３０１及びクラスタの所望の数ＮＣを使用して、重心の初期値１３０３を選択する。推奨実施例では、置換を伴わない公知のサンプリング技法を使用して、クラスタの所望の数ＮＣの各々ごとに、ベクトル集合１３０１の１つの要素を無作為に選択する。ステップ１３０４で、例えばユークリッド距離に基づく任意の公知方法を使用して、第１のクラスタ集合を生成する。その結果、複数のサンプルがＮＣ個のクラスタに分割される。ステップ１３０６で、ＮＣ個のクラスタの各々の重心を、例えばこのクラスタ内にあるベクトルの平均として、計算する。ステップ１３０８で、特異値分解論理（図７のステップ７０１）を使用して、クラスタ１３０５の固有値及び固有ベクトル１３０９を計算することができる。ステップ１３１０で、重心情報１３０７、固有ベクトル及び固有値１３０９を使用して、各クラスタごとに異なる距離関数を生成する。特定のクラスタの距離関数は、例えば、固有ベクトルによって定義された回転済み空間内のユークリッド距離を、この固有ベクトルの平方根に等しい重みで加重したものである。

【００８５】ステップ１３１２～１３１４から成るループは、新しいクラスタを生成する。ステップ１３１２で、制御論理は、ベクトル集合１３０１内の全てのベクトルにわたって、ステップ１３１３及び１３１４を反復する。ステップ１３１３で、距離関数１３１１を使用して、選択したベクトルとクラスタの重心の各々との間の距離を計算する。ステップ１３１４で、このベクトルを最も近いクラスタに割り当てることにより、クラスタ１３０５を更新する。ステップ１３１５で、終了条件に到達していれば、このプロセスを終了する。さもなければ、ステップ１３０６で、このプロセスを継続する。推奨実施例では、後続する２つの反復の間にクラスタの構成が変化していないことを条件として、このプロセスを終了する。

35

【００８６】図１４には、３次元空間における複雑な表面１４０１と、３次元の４進ツリーに基づく２つの逐次近似１４０２及び１４０３が例示されている。かかる逐次近似は、例えばH. Samet, "Region Representation Quadtree from Boundary Codes", Comm. ACM23-3, pp.163-170 (March 1980) の教示に従っている。第１の近似１４０２は、極小外接ボックスである。これに対し、第２の近似１４０３は、４進ツリー生成の第２のステップであり、そこでは各次元の中点において外接ボックスを分割するとともに、表面に交差する超方形のみを保存することによって、この外接ボックスが８個の超方形に分割されている。

【００８７】推奨実施例では、近似の階層を、ｋ次元の４進ツリーとして生成する。かかる近似の階層を生成するための本発明の方法は、例えば、クラスタの形状に対する０次近似に対応するようなクラスタ境界を生成するステップと、極小外接ボックスにより前記各クラスタの凸包を近似して、前記各クラスタの形状に対する１次近似を生成するステップと、各次元の中点において前記外接ボックスを切断することにより、前記外接ボックスを 2^k 個の超方形に区分するステップと、点を保持するような前記超方形のみを保存して、前記各クラスタの形状に対する２次近似を生成するステップと、前記保存した超方形の各々ごとに最後の２つのステップを反復して、前記各クラスタの形状に対する３次，４次，・・，ｎ次近似を逐次に生成するステップとを含んでいる。

【００８８】図１５には、クラスタの形状に対する逐次近似を使用して、所与のデータ点から一定距離よりも近い要素を保持し得るクラスタを識別するための論理の流れが例示されている。１つの実施例では、クラスタの形状は、その凸包である。他の実施例では、クラスタの形状は、全ての点を囲んでいるような連結済みの弾性表面である。この論理は、例えば図１０のステップ１００８において、候補クラスタを探索するのに使用することができる。図１５を参照して説明すると、ステップ１５０２で、形状近似の階層を有するクラスタの原集合１５０１をこのプロセスに入力するとともに、候補集合１５０５を原集合１５０１に初期化する。ステップ１５０６で、現近似を０次－形状近似に設定して、他の初期化ステップを遂行する。推奨実施例では、諸クラスタの０次－形状近似は、かかるクラスタを生成するために使用したクラスタ化アルゴリズムの決定領域によって与えられる。ステップ１５０７で、クラスタの形状の現近似とデータ点１５０３との間の距離を計算する。候補クラスタよりも遠い全てのクラスタを廃棄して、クラスタの保存集合を生成する。ステップ１５０９で、階層内に良好な近似が存在するか否かを決定する。もし、良好な近似が存在しなければ、結果集合１５１２を現保存集合１５０８に等しく設定して、このプロセスを終了する。さもなければ、ステップ１５１０で、候補集合１５０５を現保存集合１５０８に等しく設定し、現形状近似を階層内の良好な近似に設定した後、このプロセスはステップ１５０７に戻る。

36

【図面の簡単な説明】

【図１】本発明に従ったネットワーク化クライアント／サーバ・システムを例示するブロック図である。

【図２】複数データ点の分布及びクラスタ化を遂行した後の次元縮小を直感的に例示する図である。

【図３】３次元空間内の３点を２次元部分空間に射影するに際し、これらの３点のうち任意の２点間の相対距離を維持するようにした一例を示す図である。

【図４】３次元空間内の３点を２次元部分空間に射影するに際し、相対距離のランクに影響を及ぼすようにした一例を示す図である。

【図５】元の空間及び射影済みの部分空間内にある点間の距離の計算の一例を示す図である。

【図６】データベース内のデータから多次元索引を生成するための論理を例示する図である。

【図７】データの次元縮小を遂行するための論理の流れを例示する図である。

【図８】クラスタ化及び特異値分解ステップを再帰的に適用しないで生成した索引を使用して、絶対探索を遂行するための論理の流れを例示する図である。

【図９】クラスタ化及び特異値分解ステップを再帰的に適用して生成した索引を使用して、絶対探索を遂行するための論理の流れを例示する図である。

【図１０】クラスタ化及び特異値分解ステップを再帰的に適用しないで生成した索引を使用して、ｋ最近傍探索を遂行するための論理の流れを例示する図である。

【図１１】クラスタ化及び特異値分解ステップを再帰的に適用して生成した索引を使用して、ｋ最近傍探索を遂行するための論理の流れを例示する図である。

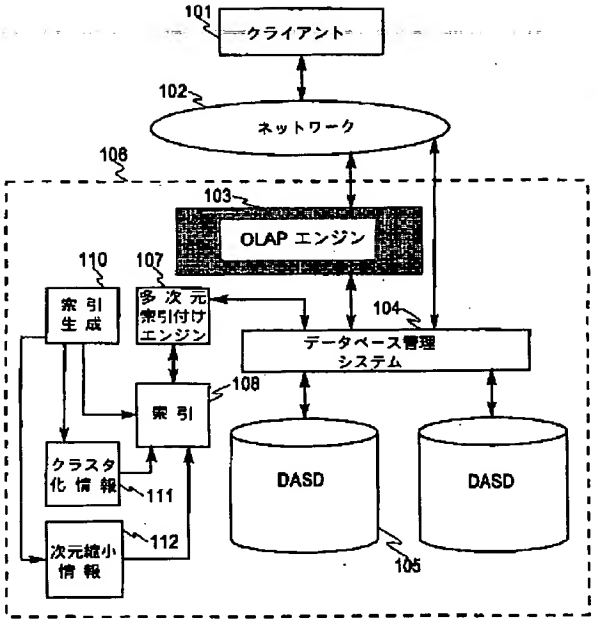
【図１２】３次元空間内にあるデータ、並びにユークリッド距離に基づくクラスタ化技法の結果及びデータの局所構造に適応させたクラスタ化技法の結果の比較を例示する図である。

【図１３】データの局所構造に適応させたクラスタ化技法の論理の流れを例示する図である。

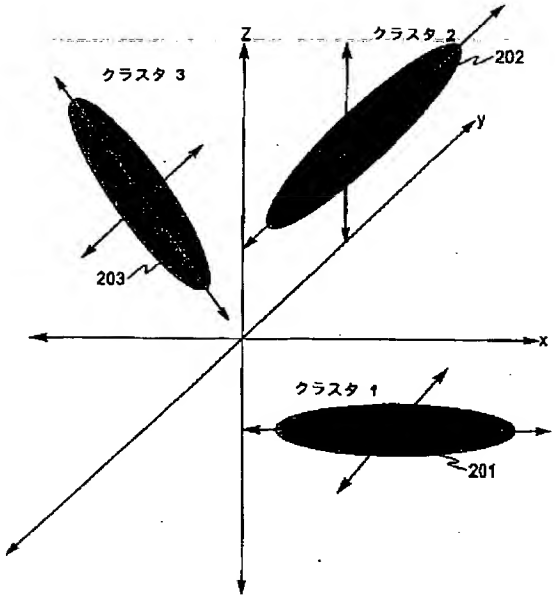
【図１４】３次元空間における複雑な超平面と、３次元４進ツリーの生成アルゴリズムを使用して生成した２つの逐次近似を例示する図である。

【図１５】クラスタの形状の逐次近似を使用して、所与のベクトルから一定距離よりも近い要素を保持し得るクラスタを決定するための論理の流れを例示する図である。

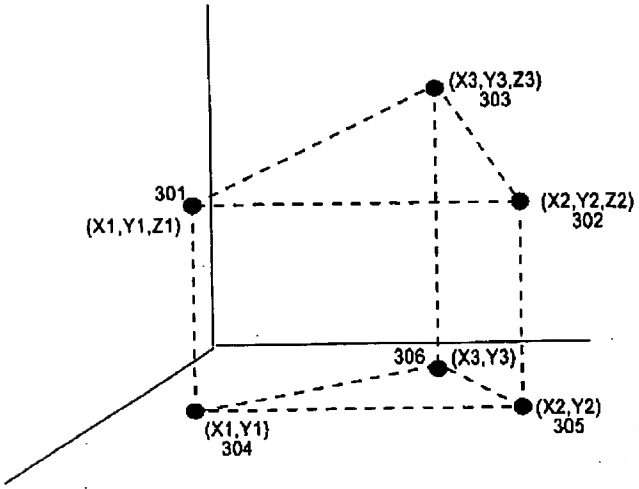
【図１】


101
クライアント
102
ネットワーク
106
103
OLAP エンジン
110
索引生成
107
多次元索引付けエンジン
104
データベース管理システム
108
索引
クラスタ化情報
111
112
次元縮小情報
DASD
DASD
105

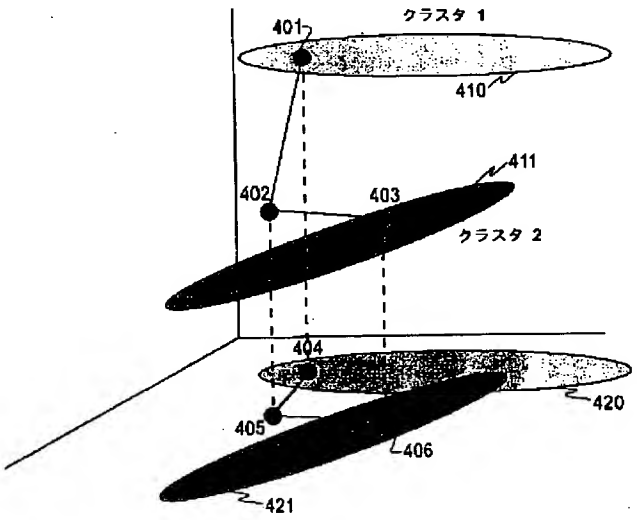
【図２】


Z
クラスタ 2
202
クラスタ 3
y
203
x
クラスタ 1
201

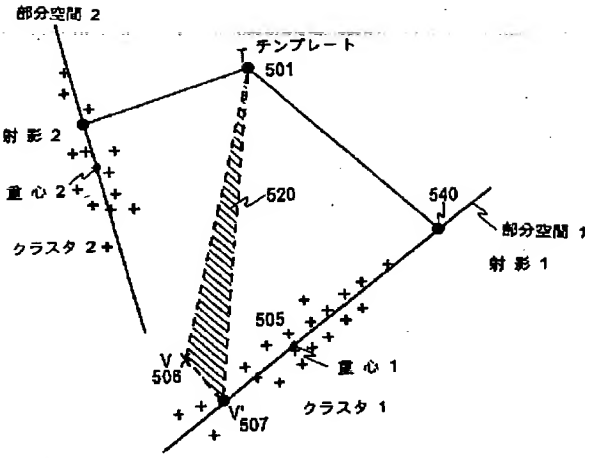
【図３】


(X3,Y3,Z3)
303
301
(X1,Y1,Z1)
(X2,Y2,Z2)
302
306
(X3,Y3)
(X2,Y2)
305
(X1,Y1)
304

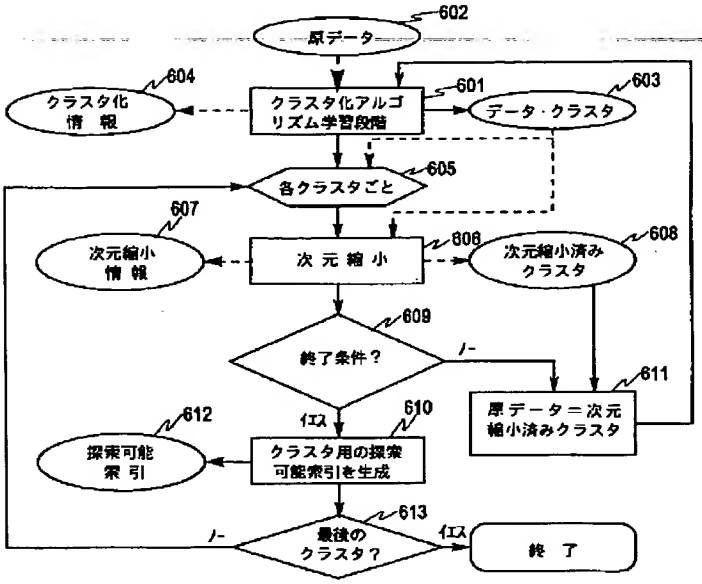
【図４】


401
クラスタ 1
410
411
402
403
クラスタ 2
404
420
405
406
421

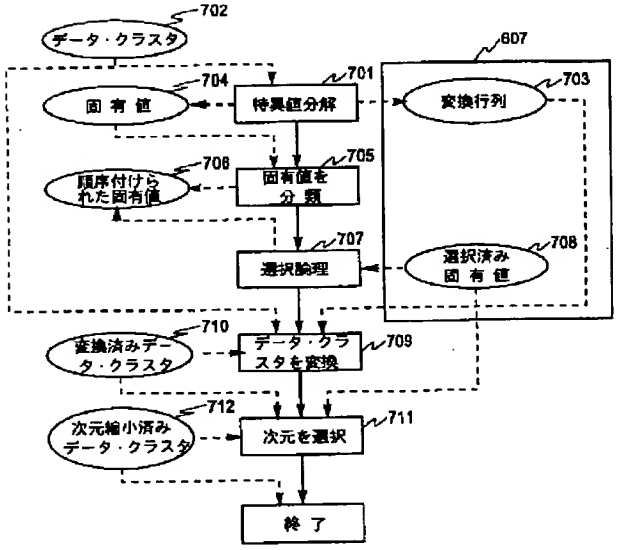
【図５】


部分空間 2
テンプレート
501
射 影 2
重 心 2
520
540
部分空間 1
射 影 1
クラスタ 2
505
重 心 1
506
507
クラスタ 1

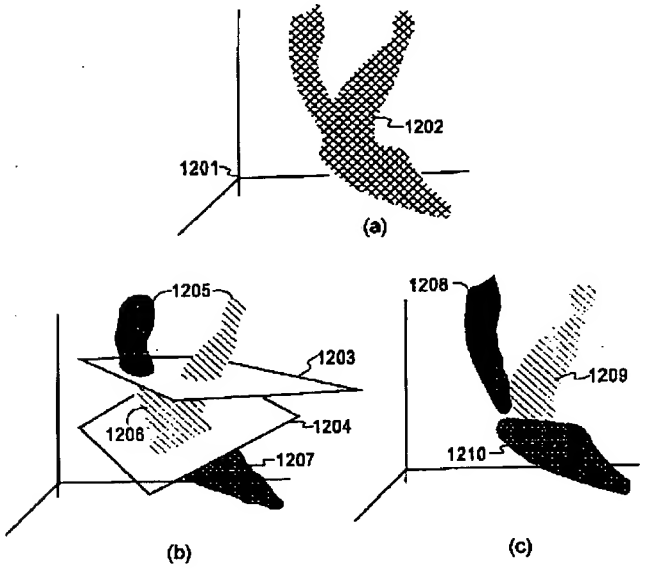
【図６】


602
原データ
604
クラスタ化 情 報
601
クラスタ化アルゴ リズム学習段階
603
データ・クラスタ
605
各クラスタごと
607
次元縮小 情 報
606
次 元 縮 小
608
次元縮小済み クラスタ
609
終了条件？
611
原データ＝次元 縮小済みクラスタ
612
探索可能 索 引
610
クラスタ用の探索 可能索引を生成
613
最後の クラスタ？
終 了

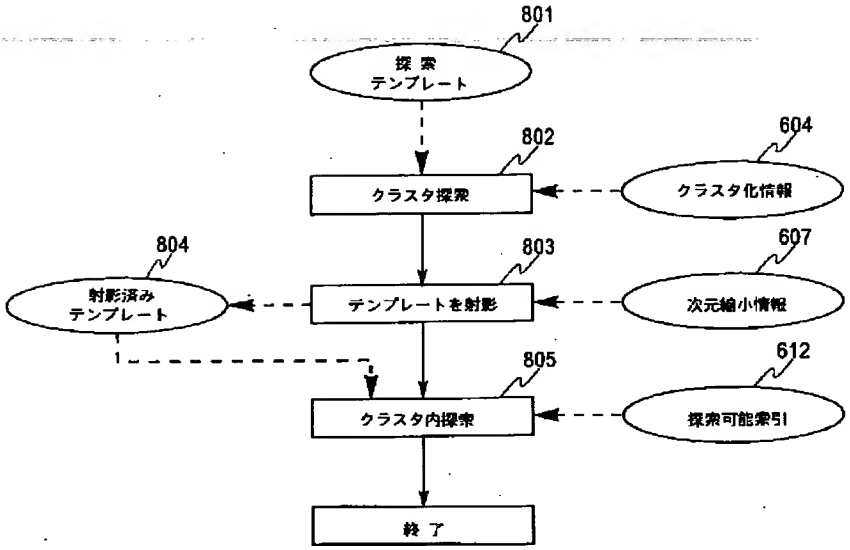
【図７】


702
データ・クラスタ
607
704
固 有 値
701
特異値分解
703
変換行列
705
708
順序付けら れた固有値
固有値を 分 類
707
選択論理
708
選択済み 固 有 値
710
変換済みデー タ・クラスタ
709
データ・クラ スタを変換
712
次元縮小済み データ・クラスタ
711
次元を選択
終 了

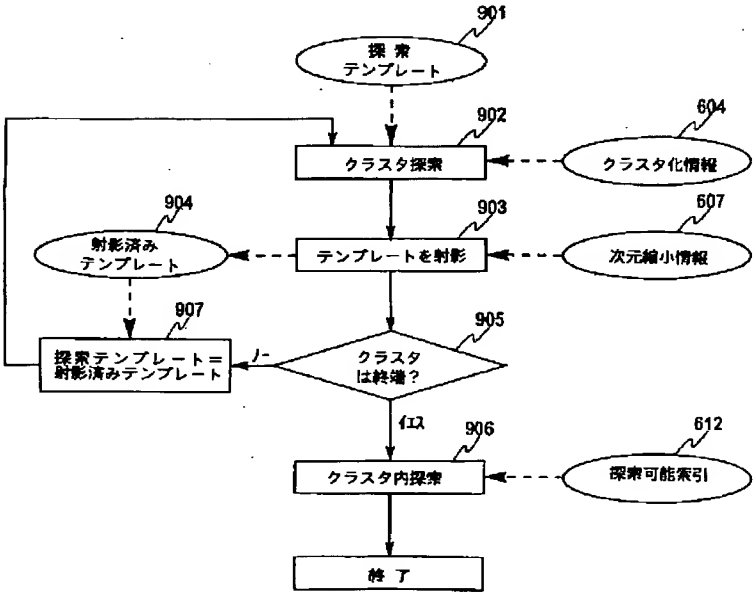
【図１２】


1202
1201
(a)
1205
1203
1206
1204
1207
(b)
1208
1209
1210
(c)

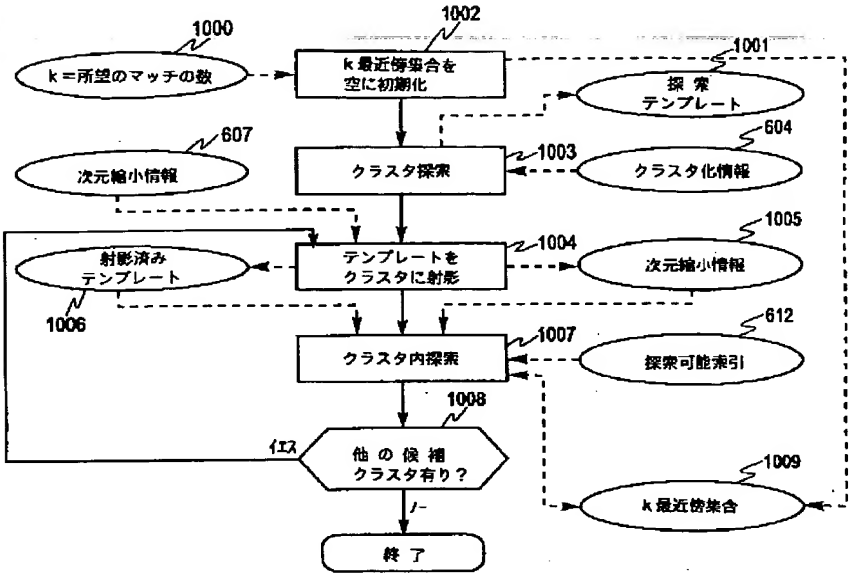
【図８】


801
探索
テンプレート
802
クラスタ探索
604
クラスタ化情報
804
射影済み
テンプレート
803
テンプレートを射影
607
次元縮小情報
805
クラスタ内探索
612
探索可能索引
終了

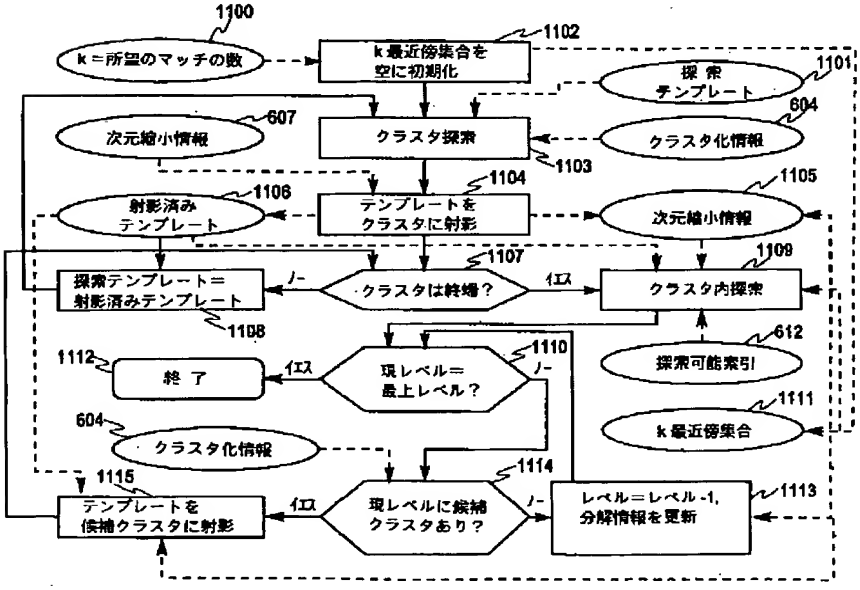
【図９】


901
探索
テンプレート
902
クラスタ探索
604
クラスタ化情報
904
射影済み
テンプレート
903
テンプレートを射影
607
次元縮小情報
907
探索テンプレート＝
射影済みテンプレート
905
クラスタ
は終端？
906
クラスタ内探索
612
探索可能索引
終了

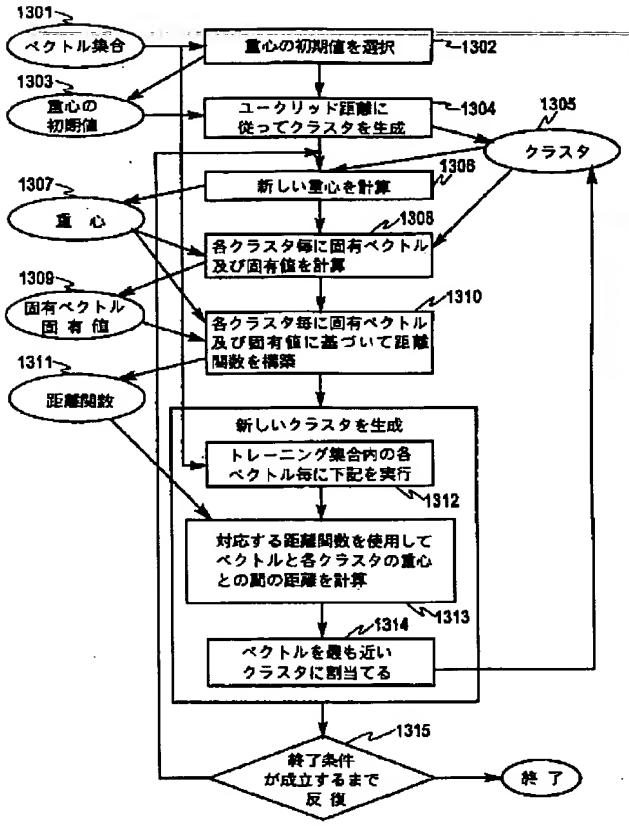
【図１０】


1000
k＝所望のマッチの数
1002
k最近傍集合を空に初期化
1001
探索テンプレート
607
次元縮小情報
1003
クラスタ探索
604
クラスタ化情報
1005
テンプレートをクラスタに射影
1004
射影済みテンプレート
1006
次元縮小情報
612
クラスタ内探索
1007
探索可能索引
1008
他の候補クラスタ有り？
イエス
ノー
1009
k最近傍集合
終了

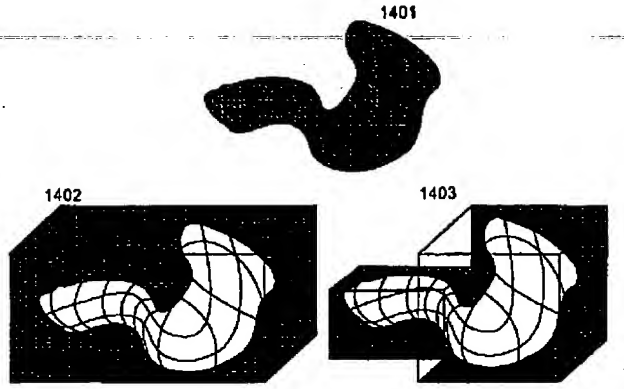
【図１１】


1100
k＝所望のマッチの数
1102
k最近傍集合を空に初期化
1101
探索テンプレート
607
次元縮小情報
1103
クラスタ探索
604
クラスタ化情報
1106
射影済みテンプレート
1104
テンプレートをクラスタに射影
1105
次元縮小情報
1108
探索テンプレート＝射影済みテンプレート
ノー
1107
クラスタは終端？
イエス
1109
クラスタ内探索
612
探索可能索引
1112
終了
イエス
1110
現レベル＝最上レベル？
ノー
1111
k最近傍集合
604
クラスタ化情報
1115
テンプレートを候補クラスタに射影
イエス
1114
現レベルに候補クラスタあり？
ノー
1113
レベル＝レベル-1，分解情報を更新

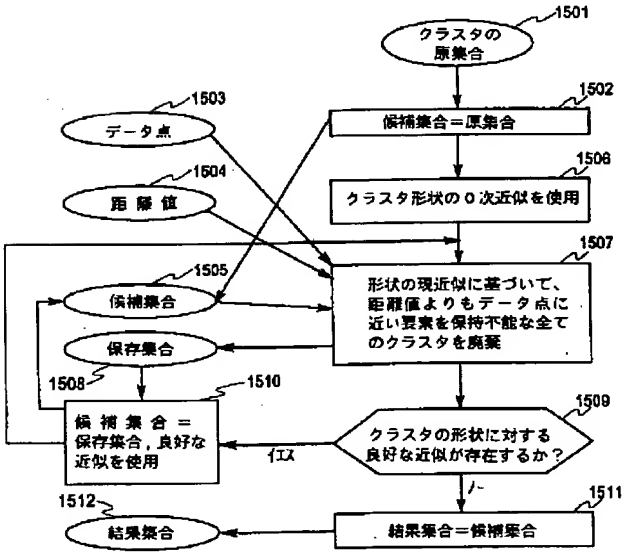
【図１３】


1301
ベクトル集合
重心の初期値を選択
1302
1303
重心の初期値
ユークリッド距離に従ってクラスタを生成
1304
1305
クラスタ
新しい重心を計算
1306
1307
重心
各クラスタ毎に固有ベクトル及び固有値を計算
1308
1309
固有ベクトル固有値
各クラスタ毎に固有ベクトル及び固有値に基づいて距離関数を構築
1310
1311
距離関数
新しいクラスタを生成
トレーニング集合内の各ベクトル毎に下記を実行
1312
対応する距離関数を使用してベクトルと各クラスタの重心との間の距離を計算
1313
1314
ベクトルを最も近いクラスタに割当てる
1315
終了条件が成立するまで反復
終了

【図１４】


1401
1402
1403

【図１５】


1501
クラスタの原集合
1503
データ点
1502
候補集合＝原集合
1504
距離値
1506
クラスタ形状の０次近似を使用
1507
1505
候補集合
形状の現近似に基づいて、距離値よりもデータ点に近い要素を保持不能な全てのクラスタを廃棄
保存集合
1508
1510
候補集合＝保存集合，良好な近似を使用
1509
クラスタの形状に対する良好な近似が存在するか？
はい
1512
結果集合
1511
結果集合＝候補集合

フロントページの続き

(72)発明者　チュンーシェン・リ
アメリカ合衆国10562、ニューヨーク州オッシニング、クロトン・アヴェニュー　50、アパートメント・２・シー
(72)発明者　アレキサンダー・トマジアン
アメリカ合衆国10570、ニューヨーク州プレザントヴィル、メドウブルック・ロード　17